The Open University

Science: a second level course

Astronomy
and
Planetary Science

Book 4

Cosmology

Book Chair: Russell Stannard

Prepared for the Course Team by
Karen Hill, Ian Robson, Russell Stannard

The Open University

S281 Course Team

Course Team Chair and General Editor Barrie Jones
Block 1 Chair Barrie Jones
Block 2 Chair Dave Rothery
Block 3 Chairs Barrie Jones and Bob Lambourne
Block 4 Chair Russell Stannard
Course Manager Cheryl Newport

Dave Adams *University of Leicester* (Author)
Jocelyn Bell Burnell (Author)
Cameron Balbirnie (BBC Producer)
Giles Clark (Publishing)
Alan Cooper (AV Production)
Sue Dobson (Graphic Artist)
Carol Forward (Course Secretary)
Peter Francis (Author)
John Greenwood (Library)
Charlie Harding (Author)
Karen Hill (Author)
Jonathan Hunt (Publishing)
Tony Jolly (BBC Series Producer)
Barrie Jones (Author)
Robert Lambourne (Author)
Jean McCloughry (Staff Tutor)
Elaine Moore (Author)
Lesley Passey (Designer)
Colin Pillinger (Author)
Ian Robson *University of Central Lancashire* (Author)
Dave Rothery (Author)
Dick Sharp (Editor)
Russell Stannard (Author)
Liz Swinbank (Consultant)
Margaret Swithenby (Editor)
Arnold Wolfendale *University of Durham* (Course Assessor)
Ian Wright (Author)
John Zarnecki *University of Kent* (Author)

Cover: All-sky map showing departures from uniformity in the cosmic background radiation. Courtesy of NASA Goddard Space Flight Center, Greenbelt, Maryland.

The Open University, Walton Hall, Milton Keynes, MK7 6AA.
First published 1994

Edited, designed and typeset by The Open University.
Printed in Great Britain by Henry Ling Ltd, The Dorset Press, Dorchester, DT1 1HD.
ISBN 0 7492 5128X
This text forms part of an Open University Second Level Course. If you would like a copy of *Studying with The Open University*, please write to the Central Enquiry Service, PO Box 200, The Open University, Walton Hall, Milton Keynes, MK7 6YZ. If you have not enrolled on the Course and would like to buy this or other Open University material, please write to Open University Educational Enterprises Ltd, 12 Cofferidge Close, Stony Stratford, Milton Keynes, MK11 1BY, United Kingdom.
1.1

Cosmology

Contents

Introduction and study guide for Block 4

Cosmology is the study of the Universe. A rather esoteric subject perhaps, yet one that deals with truly fascinating questions – the really big issues that have puzzled humankind down through the ages. For example, is space infinite? Did the Universe have a beginning? Will it have an end? Questions don't come much bigger than this!

Traditionally, the search for answers was thought to be a matter for speculation only. Indeed, at the turn of the 19th century, cosmology hardly ranked as a science at all. There is only one Universe, so repeated control experiments under different conditions were out of the question. And in any case, the Universe was not exactly something one could actively experiment with.

Yet despite these drawbacks, in recent years remarkable advances have been made. This has been thanks to new observations, and to the theoretical framework provided by Einstein's general theory of relativity. We now seem to be much closer to answering some of these age-long questions.

The book consists of two chapters. The first is concerned with presenting the experimental evidence for believing the Universe began with a Big Bang. The second takes a closer look at the nature of the expansion of the Universe in the light of Einstein's theory. In this latter chapter, we introduce you to a very new and exciting idea called the inflationary theory. It arises out of the study of high-energy particle physics – the examination of the smallest constituents of nature. It comes as something of a surprise to learn that the key to understanding the past and the future of the Universe as a whole lies in a better understanding of the behaviour of its smallest parts.

There is a television programme – TV programme 8, *Cosmology on trial* – associated with the Block; this reviews the evidence for the Big Bang. There are also two video sequences – video sequence 13, *The expanding Universe*, and video sequence 14, *The geometry of the Universe*. The first will help you to understand better the type of expansion involved in the Big Bang; the second demonstrates physical models through which we might visualize the various forms of geometry permitted by the theory of relativity. Finally, audio band 4, *Creating a short talk*, provides some guidance as to how to tackle part of the TMA associated with the Block – an assignment in which you are invited to prepare a short talk for a general audience.

After you have finished your study of Book 4, you should view video sequence 15, *Detecting radiation*. This brings together some of the radiation detectors introduced in the earlier audiovisual material.

It is important that you read the associated notes before you watch the TV programme or video sequences, or listen to the audio band.

The authors of this Block trust that you will find it a very fitting conclusion to your study of the Course.

Chapter 1
The expanding Universe

Prepared for the Course Team by Karen Hill, Ian Robson and Russell Stannard

1.1 Introduction

At the beginning of the 20th century, **cosmology** – the study of the composition, behaviour, and development of the Universe as a whole – received only limited scientific attention. The problem was that there simply did not exist an adequate observational and theoretical foundation on which to build a consistent picture of the whole Universe. There were serious attempts to determine the origin of the Solar System but the concept of the overall Universe was limited. To the extent that it incorporated any science, the thinking was based on Newtonian theory. From his work on the orbits of the Moon and the planets in the Solar System, Newton realized that gravity was a long-range force. In the absence of any other force, it would come to dominate the large-scale interaction of bodies in the Universe. Therefore, all the bodies in the Universe would attract each other, and so ought to end up as a single clump of material. This was obviously not the case. Newton puzzled over why this was so. One way out was to introduce an additional force – a repulsive one this time – to counter gravity and hence stave off collapse. But there was no evidence for such a force from Newton's work. Neither has more recent work yielded any consistent evidence for a long-range repulsive force. An alternative was to postulate that the Universe was infinite in extent. With all the matter nicely arranged in a homogeneous distribution, the mutual attraction on each body by its partners would cancel out, and a stable configuration could be expected. Although an infinite universe sounded somewhat contrived, this was the preferred solution.

Little further happened to the subject until Einstein developed his general theory of relativity over the period 1905–15. As we shall be seeing in the next chapter, this theory replaced Newton's earlier one, and was destined to become the main theoretical framework for tackling cosmological questions.

The equations of the theory pointed in a natural way to the conclusion that the Universe was a dynamic one. Einstein could have *predicted* that the Universe was either expanding or contracting! Unfortunately, like everyone else, he was under the influence of the prevailing view of the time that the Universe was static. Thus, in order to bring his theory into line with what he assumed was 'observation', he added a constant to his equations – one that represented a hypothetical repulsive force. In the light of the subsequent work of Hubble, showing that the Universe is expanding, Einstein was later to describe the introduction of his spurious constant as 'the greatest blunder of my life'.

No less important in transforming cosmology into a serious science, was the development of experimental techniques: powerful telescopes that were able to probe to great depths, and equipment for detecting and analysing electromagnetic signals both within and beyond the visible range.

In this chapter, we shall set before you the relevant information that has been gathered by these various experimental techniques, and show how they combine to produce convincing evidence that the Universe began with a Big Bang.

1.2 Olbers' paradox

But before we launch into modern cosmology, let us remind ourselves of one of the oldest puzzles posed by philosophers, one that is based on an observation we have all repeatedly made – regardless of whether or not we happen to possess a telescope: *Why is the sky dark at night?*

Surely the answer is obvious: 'The Sun has set and therefore the sky becomes dark'. But what about the stars? They are suns too. Imagine a universe

populated by stars. Let us assume it is static and infinite in extent, and that the stars are distributed isotropically (the same number in all directions) and homogeneously (the same number in all volumes). In this case, when we look out into the night sky we would expect to see a star in whichever direction we looked; eventually the line of sight would intercept a star as shown in Figure 1.1. Because all stars are luminous bodies like the Sun, then the night sky should be uniformly as bright as the surface of a typical star, like the Sun.

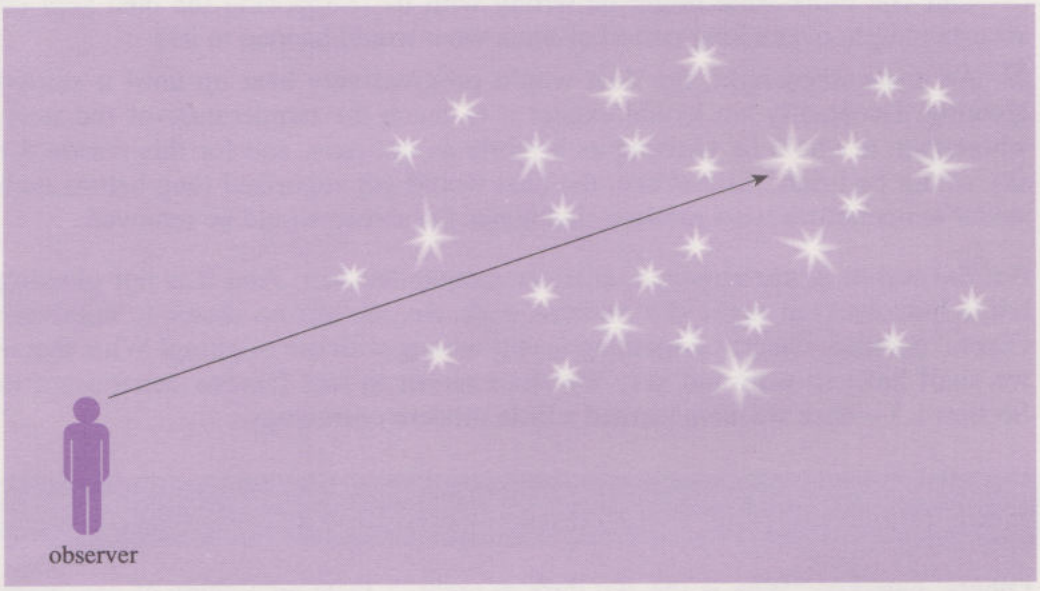

Figure 1.1 In a static universe of infinite extent, where the distribution of stars is isotropic and homogeneous, any line of sight drawn from an observer must eventually end at the surface of a star.

Clearly this is not the case; the night sky is black (at least, it is once one is away from town lights!). Therefore there must be something wrong with either our assumptions or our reasoning. This was the famous paradox posed in 1826 by the Viennese astronomer Heinrich Olbers (1758–1840). Although others before him had also commented on the puzzle, it is now known as **Olbers' paradox**.

But wait a moment, you might say. What about the fact that the flux density we receive from the star falls off as the inverse of the distance (as we saw in Book 1, Subsection 2.3.4)? Surely the light we receive from a more distant star would be weaker than that from a nearer one. This is perfectly true – but it does *not* explain the paradox. Consider a star at distance, d_1, from the Earth. The flux density we receive from it is $\propto 1/d_1^2$. Similarly, from an identical star at d_2, the flux density received is $\propto 1/d_2^2$. If $d_2 > d_1$, obviously the flux density from the further star at d_2 is less than that from the star at d_1. But consider our assumption of homogeneity – the spatial density of stars in any volume being the same. The number of stars in a shell of radius d_1 and thickness Δd, centred on the position of the observer (Figure 1.2), is proportional to its volume, i.e. $4\pi d_1^2 \Delta d$. Thus the light received from the stars in this shell is $\propto 4\pi d_1^2 \Delta d \times 1/d_1^2$, which is $\propto 4\pi\Delta d$. Likewise, the number of stars in a shell of radius d_2 and thickness Δd will be proportional to its volume, $4\pi d_2^2 \Delta d$, and the light received from the stars is $\propto 4\pi d_2^2 \Delta d \times 1/d_2^2$, which is again $\propto 4\pi\Delta d$.

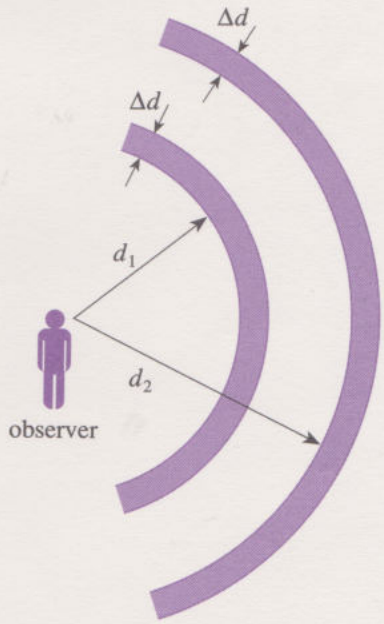

Figure 1.2 Two shells of radius d_1 and d_2, and thickness Δd, drawn about the position of the observer.

☐ What do you conclude from this?

■ The total flux density received by the observer from the stars in a shell is independent of the radius of the shell. It does *not* fall off with distance. Each star at the larger distance delivers a smaller flux density to the observer, but this is compensated by there being more stars at the larger distance.

A second objection you might wish to raise is that the medium through which the light has to pass could be dusty. The more dust the light has to pass through, the more likely it is to be absorbed. Hence light from the furthermost stars would be the most dimmed. On the face of it, a very plausible argument.

□ Can you think what might be wrong with it? (Supposing the dust kept on absorbing light over a long period of time, what would happen to it?)

■ As it absorbed light, the dust would progressively heat up until it started glowing. Eventually we would expect it to reach the temperature of the stars, whereupon it would be glowing as brightly as the stars, and for this reason the sky would be bright. In practice, the dust would get vaporized long before such stellar temperatures were reached, and hence the screen would be removed.

But the screen is *not* removed; there *is* interstellar dust. And it is not glowing white hot. And the night sky remains dark. So, we are no closer to resolving Olbers' paradox. Clearly something is very wrong with our thinking! What that is we shall have to wait and see. We shall return to this famous conundrum in Section 1.5 – once we have learned a little modern cosmology.

Summary of Section 1.2

Olbers' paradox, 'Why is the sky dark at night?', rests on a chain of argument that appears to indicate that the sky should be bright:

1. In an infinite, isotropic and homogeneous universe, any line of sight drawn from an observer must eventually end at the surface of a star.
2. Flux density received from a star decreases with distance d ($\propto 1/d^2$), but the number of stars in a shell centred on us increases with distance ($\propto d^2$), so the total flux density received from any shell does *not* fall with distance.
3. Absorption of light by intervening dust cannot be the answer; the dust would glow white hot and then be vaporized.

1.3 Hubble's law and the expansion of the Universe

1.3.1 Origin of the redshift

We met **Hubble's law** in Book 3, Subsection 2.3.5:

$$z = \frac{H_0}{c}d \tag{1.1}$$

where z is the redshift of the light received from a galaxy, distance, d, from us. (The subscript on Hubble's constant, H_0, reminds us that this is the value of the constant as it applies to the present epoch, t_0; at earlier times, its value would have been different.) You also learnt that redshift z is proportional to the difference between the wavelength of a spectral line as observed by ourselves and its value at emission, and is defined by

$$z = \frac{\lambda_{obs} - \lambda_{em}}{\lambda_{em}} \tag{1.2}$$

where λ_{obs} is the observed wavelength, and λ_{em} is the wavelength at the point of emission.

So far we have said nothing about the origin of this redshift. What we do know is that if a light source is receding from us, the wavelength of the light we receive from it will be stretched out. We say it is Doppler-shifted. So, is this the explanation of the Hubble redshift? Are the galaxies receding from us, and is the light Doppler-shifted? The answer is yes – and no! The galaxies *are* receding from us; the Universe is expanding. And this expansion is the origin of the redshift. However, it is not really correct to say that the light is Doppler-shifted – not in the normal way. That would be to adopt a Newtonian view of the motion: the only motion would be that of the galaxies rushing outward into dark, empty, infinite space. Such a picture is inadequate. To understand why, we shall have to wait until Chapter 2. Meanwhile we note that for *small* redshifts ($\lesssim 0.1$), the Newtonian viewpoint offers useful results, in that it gives the correct value for the rate of increase of distance, and that is the model we shall adopt for the time being.

The Doppler shift for a light source receding from us at a speed v which is very much smaller than c is

$$z = v/c \tag{1.3}$$

ITQ 1.1 We first met the Doppler shift in Book 1, Chapter 2 (Equation 2.3), where it was defined in terms of frequencies. Use this equation together with Equation 1.2 to derive Equation 1.3.

From Equations 1.1 and 1.3 we therefore obtain an interpretation of Hubble's law for *small* redshifts:

$$v = H_0 d \tag{1.4}$$

where the recessional speed v of the galaxy is usually expressed in $\mathrm{km\,s^{-1}}$, the distance d in Mpc, and Hubble's constant has units of $\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

That this interpretation is valid only for *small* redshifts arises from the equation on p. 58 of Book 1 applying only when $v \ll c$.

Figure 1.3 illustrates this interpretation. It shows some experimental data of z versus d plotted as v versus d using Equation 1.4. We see that it follows a straight-line relationship.

Thus, the distances between the galaxies and our Galaxy are increasing at a rate that is proportional to their distances from us.

ITQ 1.2 (a) Using a value for Hubble's constant of $75\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, calculate the distance in Mpc to a quasar of redshift 0.12. (b) Calculate how rapidly this distance is increasing.

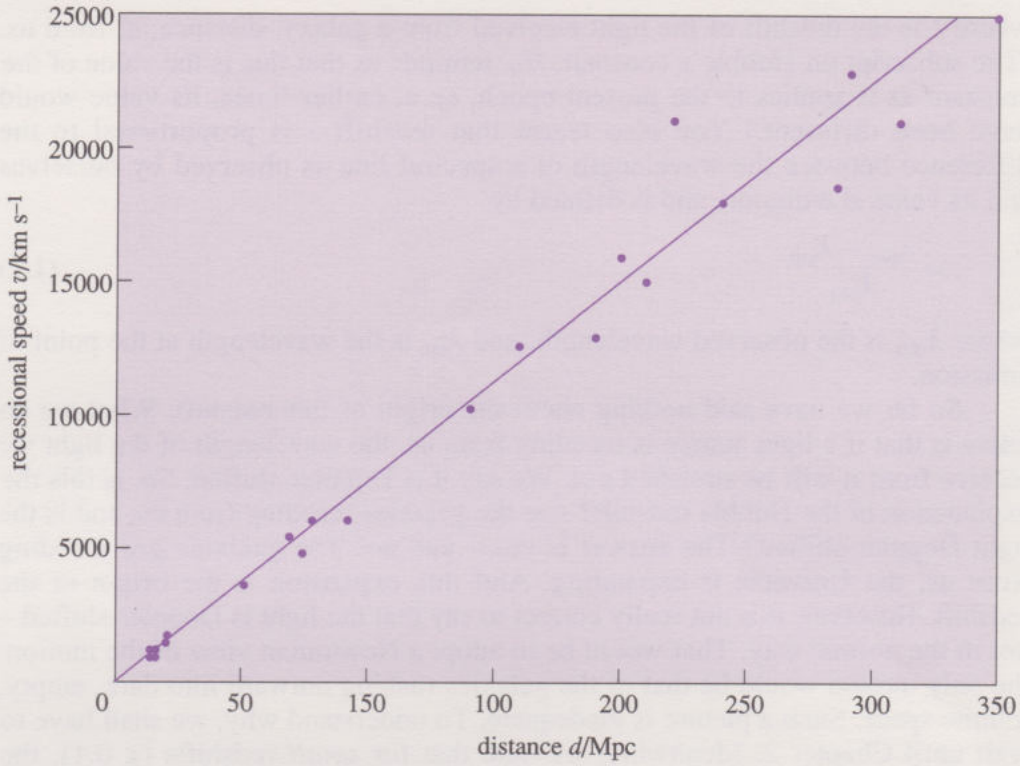

Figure 1.3 The recessional speed of distant galaxies as a function of their distance from us.

1.3.2 The cosmological principle

The fact that the distances between us and the galaxies are increasing appears at first sight very odd. The simplest interpretation would be that we lie at the centre of the Universe! Yet everything we have learned about ourselves so far in this Course points to us *not* having claims to being anything special; we live on an ordinary (albeit precious) planet, going round a common-or-garden star, in a medium-sized spiral galaxy of no repute.

This is a good point to introduce a fundamental tenet of cosmological studies, one which we have used already but without drawing any special attention to it. It raises our very ordinariness to a statement of principle: the **cosmological principle**. According to this, it becomes an article of faith that the Universe is everywhere **homogeneous** and **isotropic**. Homogeneous means that it has the same density everywhere. If this is the case, then it will also look the same in all directions – which is what isotropic means. We note that we must take sufficiently large volumes of space for this to hold, i.e. for local irregularities to have ironed themselves out. Clearly within the Solar System and the Galaxy it fails severely. Indeed, even on the scale of clusters of galaxies the principle does not hold too well (this is discussed in Chapter 4 of Book 3, and in video sequence 11). Nevertheless, despite these difficulties, we *do* adopt the cosmological principle. Without it, we would find it difficult to establish any kind of theoretical framework for handling cosmology.

Note that the converse does *not* hold – an isotropic universe need not be homogeneous.

The cosmological principle leads to the important conclusion that no-one – including ourselves – can be thought of as occupying a privileged position. Any position in space is as good as any other. In other words, no-one can regard themselves as exclusively located at the centre of the Universe.

An immediate consequence of this is that observers in other galaxies must make the same kind of observations as we do. If we see the distance between us and the other galaxies increasing with a rate that is proportional to their distance from us, observers in other galaxies must see the same kind of thing relative to themselves. It is very important that you grasp this idea. In Figure 1.4a, various galaxies are labelled 1, 2, 3 and 1´, 2´, 3´. They are conveniently situated with equal separations *d*. The galaxies 1 and 1´ are receding from us (the Milky Way, labelled 0), at the centre of the picture, with speed v. Those at twice the distance, 2 and 2´, are receding at speed $2v$, and so on, in accordance with Hubble's law.

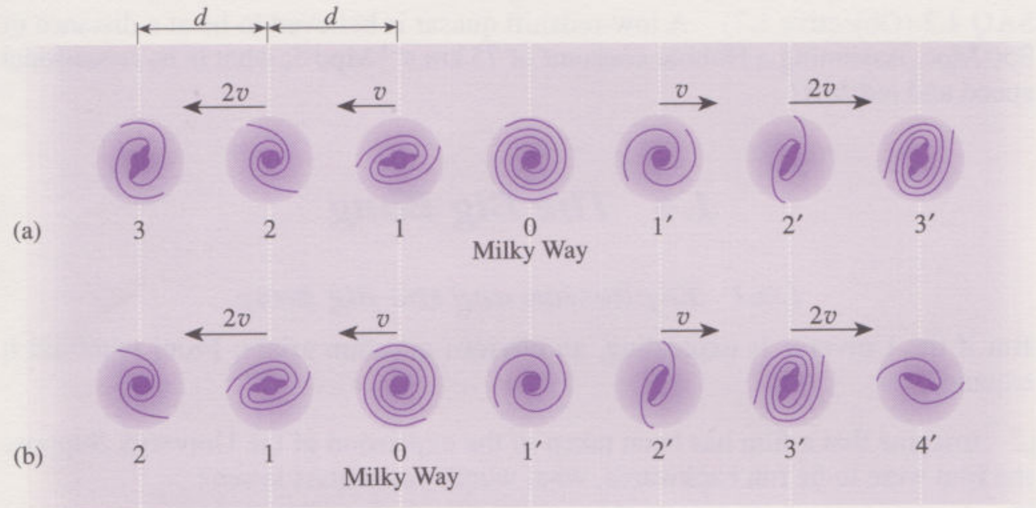

Figure 1.4 (a) Schematic drawing showing a number of equally spaced galaxies viewed from the Milky Way at the centre. (b) Here the viewpoint has been transferred to galaxy 1´.

□ What observations would observers on galaxy 1´ make of the apparent speeds of ourselves and of the other galaxies?

■ Since galaxy 1´ is moving to the right (in the positive direction) with a speed v, i.e. at a velocity $+v$, relative to us, we must subtract velocity $+v$ from each of the velocities in Figure 1.4a to find the motion of the galaxies relative to 1´. We obtain Figure 1.4b. From this we conclude observers on galaxy 1´ see the same pattern of motion, the same Hubble law, as we do.

This is in accord with the cosmological principle. Thus, we arrive at the conclusion that, for the cosmological principle to hold, if the galaxies are receding then the rate of increase of distance of the galaxies from us must be proportional to their distance from us. The linearity of Hubble's law can therefore be taken as supporting evidence for the cosmological principle, even though the data are taken from a single viewpoint. Not only is the Universe expanding from our point of view here in the Milky Way, it will then be expanding in the same way regardless of the viewpoint adopted.

Summary of Section 1.3 and SAQs

1. Hubble's law, $z = H_0 d/c$, relates the redshift, z, of spectra observed by ourselves to the spectra emitted by the source at a distance *d*.
2. Only for *small* redshifts, z, is it useful to interpret the redshift of galaxies as Doppler shifts. In this interpretation, $z = v/c$, and distance is then related to the recessional speed v, by $v = H_0 d$.

3 The cosmological principle states that the Universe is everywhere homogeneous and isotropic.

4 It follows from the cosmological principle that (at small redshifts) the recessional speeds of the galaxies are proportional to their distance from the observer. This is in accordance with observation (Hubble's law).

SAQ 1.1 (Objective 1.3) For the quasar mentioned in ITQ 1.2, calculate the wavelength at which the ultraviolet Lyman α line of wavelength 121.6 nm is observed, and say in which part of the spectrum it lies.

SAQ 1.2 (Objective 1.3) A low-redshift quasar is believed to lie at a distance of 350 Mpc. Assuming a Hubble constant of 75 km s^{-1} Mpc^{-1}, what is its recessional speed and redshift?

1.4 The Big Bang

1.4.1 Expansion and the Big Bang

But if the Universe is expanding, an obvious question arises: From what did it expand?

□ Imagine that a film has been taken of the expansion of the Universe. Suppose the film were to be run backwards, what would you expect to see?

■ Obviously the galaxies would appear to move towards each other. But more than that: a galaxy that was twice as far away as another would appear to move twice as fast; one that was three times further away would appear to move three times as fast. The net result would be that, not only would all the matter of the Universe converge, it would all arrive together *at the same time*.

In other words, thinking now of the film being run the correct way, it seems reasonable to suppose that all the matter we see around us began at some point in the past from a condition of infinite density, and that it then expanded. The assumption that this was the case is a *central* feature of the **Big Bang** theory.

At this juncture, it will be useful to introduce a **scale factor**, $\boldsymbol{R(t)}$, that tells us how the Universe scales up with time. $R(t)$ is the function of time by which all distances scale in a universe that satisfies the cosmological principle. For example if $d(t_0)$ is the distance between two galaxies now at **cosmic time**, t_0, then the distance, $d(t)$, at some other cosmic time, t, is given by

Cosmic time is the time since the beginning of the expansion of the Universe. $t = 0$ is the instant of the Big Bang. The cosmic time now, t_0, is the age of the Universe.

$$d(t) = \frac{R(t)}{R(t_0)} \times d(t_0) \qquad (1.5)$$

The significance of the scale factor is that it describes only a simple scaling up or scaling down of distances as a function of time. This is consistent with, and required by, the cosmological principle. It does not allow, for example, for rotations or shears in the Universe – these would enable us to distinguish one location from another. (The absence of rotations and shears is something which must be tested observationally to see whether the cosmological principle is violated.)

As cosmic time, t, tends to zero and we extrapolate backwards towards the Big Bang at $t = 0$, $R(t)$ also tends to zero and any finite region of the Universe shrinks towards a point. Occupying no volume, the matter has infinite density. This is not something that physicists are happy with; they are unable to deal with such infinite quantities. The suspicion therefore is that as we imagine going back in time towards the instant of the Big Bang, a stage will be reached where

physics as we know it will break down. In Chapter 2, you will be able to see that physicists are able to interpret and investigate events exceedingly close to the instant of the Big Bang. Yet always there remains that tantalizing first tiny fraction of a second where it appears we have to throw in the sponge.

1.4.2 The age of the Universe

But returning to the expanding Universe, Figure 1.5 shows a plot of $R(t)$ versus t, with the current epoch being labelled by t_0. This simple picture, obtained through a combination of Hubble's law and the cosmological principle, is the first strong pointer to a universe that came into being at some finite time in the past, and subsequently expanded at the more-or-less constant rate we see today. When did this occur? It is very tempting, as is shown on Figure 1.5, to extrapolate backwards until $R(t)$ becomes zero.

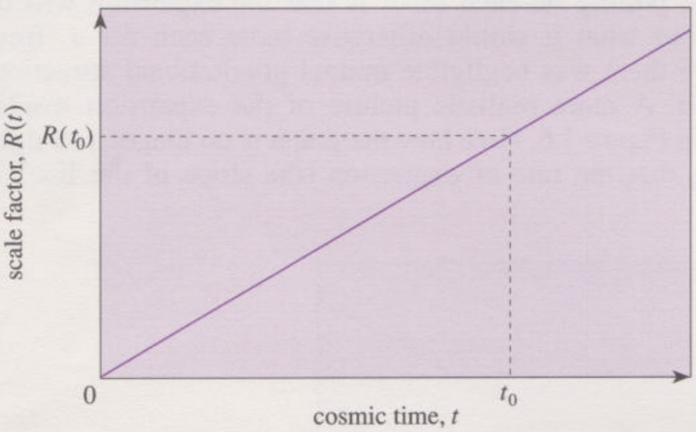

Figure 1.5 The scale factor, $R(t)$, plotted as a function of cosmic time, t.

The gradient of the line in Figure 1.5 (namely $R(t_0)/t_0$, where $R(t_0)$ is the value of $R(t)$ at the current epoch) is the rate at which the scale of the Universe is increasing. In Figure 1.5 we assume that the gradient is constant (we will later revise this).

If the Universe had really expanded at a constant rate as in Figure 1.5, then a galaxy at a distance d would have been receding from us at a constant speed given by

$$v = \frac{d}{t_0} \tag{1.6}$$

But v, for small redshifts (assume the galaxy is not far off), is given by the form of Hubble's law in Equation 1.4. Thus, eliminating v between Equations 1.4 and 1.6, we get

$$d/t_0 = H_0 d \tag{1.7}$$

and so

$$t_0 = 1/H_0 \tag{1.8}$$

This highlights that H_0 is a quantity of type 1/time, apparent also from the units $\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

If we believe Figure 1.5, then the time the Universe has taken to expand from the instant of the Big Bang to the current epoch is t_0. In order to measure the age of the Universe, therefore, all we have to do is to obtain the value of Hubble's constant and the inverse tells us the age. But remember, as we saw in Book 3, Subsection 2.3.5, there is uncertainty in the value of H_0 by about a factor of two, the agreed value lying somewhere between 50 and 100 $\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

ITQ 1.3 (a) Choosing a value for H_0 of 50 km s^{-1} Mpc^{-1}, estimate the age of the Universe in years. (b) Repeat the exercise with H_0 = 100 km s^{-1} Mpc^{-1}.

Note that even though the discussion here has been restricted to small redshifts (we used Equation 1.4), and hence to *nearby* galaxies, considerations of age based on when the matter in these particular galaxies occupied a negligible volume, must apply to the *whole* Universe.

These estimates of the age of the Universe are based on the very tempting assumption that we are justified in extrapolating backwards in time to $t = 0$. This in turn depends on Figure 1.5 being a realistic representation of the way the Universe has evolved.

In reality, Figure 1.5 may be too simplistic. Let's consider the reasons why. Every object in the Universe has mass, and this mass produces a gravitational attraction which will act on all the other matter in the Universe. The effect of all the parts of the Universe pulling on each other is that the expansion will be slowed down compared to what it would otherwise have been for a 'free' expansion, i.e. one where there was negligible mutual gravitational attraction, and hence no retardation. A more realistic picture of the expansion would, therefore, be that shown in Figure 1.6. Note how the graph is no longer a straight line; it is one indicating that the rate of expansion (the slope of the line) is reducing with time.

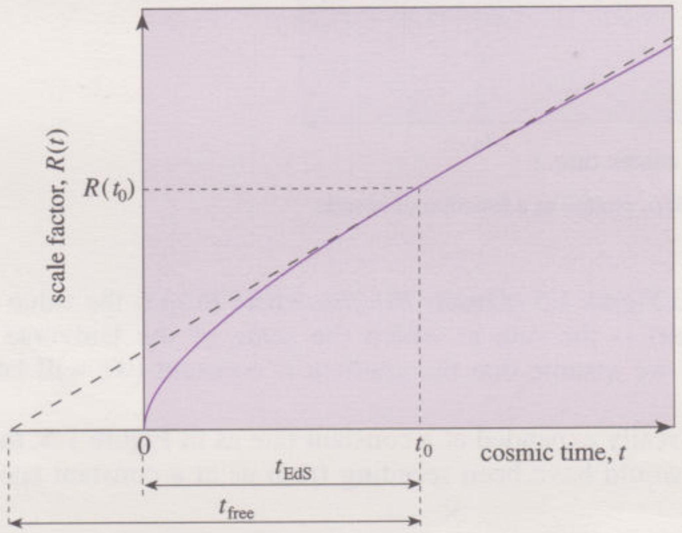

Figure 1.6 A plot of scale factor, $R(t)$, versus t for the Einstein–de Sitter model of the Universe.

In order to account for the fact that the curve deviates from a straight line, a new constant has to be introduced into the theory to augment Hubble's constant. This second constant is called the **deceleration parameter**, $\boldsymbol{q}$. It is a measure of the rate of change of the slope with time.

1.4.3 The Einstein–de Sitter model

The particular curve we have drawn in Figure 1.6 is a rather special one: at very large times (well beyond the scale drawn in Figure 1.6) its slope approaches zero, i.e. it becomes parallel to the t axis. This is the case in the **Einstein–de Sitter model** of the Universe (so-called after Albert Einstein (1879–1955) and the Dutch astronomer Willem de Sitter (1872–1935), who both advocated its use). It is one where the kinetic energy of the Big Bang and the gravitational energy of the Universe are chosen to be so finely balanced as to produce this long-term

behaviour. It is a particularly useful model in that it facilitates various calculations. It should be stressed, however, that this curve is only one of a possible series of curves that we shall be exploring in Chapter 2.

Considering Figure 1.6, were it to be extrapolated, then at extremely large values of cosmic time, the scale factor no longer changes, i.e. the expansion comes to a halt – the galaxies cease to recede. Under those conditions we would be living in what, to all intents and purposes, would be a static universe.

1.4.4 A second look at the age of the Universe

Looking again at Figure 1.6, we can note that according to this model – and indeed according to *any* model that allows for a measure of self-gravity – the value we quoted earlier for the age of the Universe is an overestimate. That was the value obtained by taking the slope of the curve at the present epoch, and seeing where it intersects with the time axis. The dashed line in Figure 1.6 is, in fact, the previous plot we had in Figure 1.5 – the one appropriate to free expansion. The value, t_{free}, shown in the figure is our original estimate of cosmic time, t_0, obtained from $t_{\text{free}} = t_0 = 1/H_0$. But what we now see is that the age of the Universe for the Einstein–de Sitter model would be the smaller value, $t_{\text{EdS}} = t_0$.

Self-gravity: all matter distributed in a volume will gravitationally attract all other matter in that volume. This will lead to a slowing down of the expansion.

For this particular model, the relationship in Figure 1.6 between $R(t)$ and t can be shown to have the form:

$$R(t)/R(t_0) = (t/t_0)^{2/3} \tag{1.9}$$

From this, with a little help from simple calculus (beyond the scope of this Course), it can further be demonstrated that

$$t_{\text{EdS}} = 2/(3H_0) \tag{1.10}$$

Remember that H_0 is the value of Hubble's constant *today*, at t_0.

$$t_{\text{Eds}} = \left(\tfrac{2}{3}\right)t_{\text{free}} \tag{1.11}$$

1.4.5 Which model describes the Universe?

So, which kind of model best describes the behaviour of the Universe? What do observational tests have to say?

We can begin by focusing on what we might deduce about the age of the Universe from what we have already learnt about the ages of various objects.

In Book 2, Section 2.1, we saw that the age of the Earth is estimated to be about 4.5 billion years. We can, therefore, be confident that the Sun is at least that old. However, the Sun is a relatively late arrival on the cosmic scene, so that means the Universe must be significantly older.

What about other stars? The oldest stars in the Milky Way are the population II stars in globular clusters. We saw in Book 3, Subsection 1.4.2, that there is uncertainty as to the exact ages of these stars, but we believe them to lie in the range 10–17 billion years. Thus, the Universe must be older than this. It is this estimate of age which begins to put severe constraints on the age of the Universe and hence the form of the curve in Figure 1.6. It also has implications for the value of Hubble's constant. If globular cluster stars are at the upper end of their age estimate, around 17 billion years, then we can say that if the evolution of the Universe follows the Einstein–de Sitter model of Figure 1.6 (or one with more self-gravity), then we cannot have a high value of Hubble's constant; it must be around 40 km s^{-1} Mpc^{-1} or less.

But is there no other way of choosing between the various possible cosmological models? Why not just measure the form of Figure 1.6 directly by

observation? Remember that by measuring distances of extremely distant objects we are in fact looking at epochs in the past. The finite speed of light means that we are observing the Universe as it was when the light was emitted from these distant objects, not as it is now. So in principle all we need to do is measure the scale factor, $R(t)$, at a variety of distances, and this will give us $R(t)$ as it was at the corresponding epochs. Knowing the variation of $R(t)$ with cosmic time, we could then find the value of the deceleration parameter, q. For a freely expanding universe, q would be zero; for an Einstein–de Sitter universe it turns out to have a value of $\frac{1}{2}$. Other models are characterized by other values for q.

At this point you might want to remind yourself of the content of Section 1.5 of *Preparatory science*.

One way to measure $R(t)$ is to measure redshifts at greater and greater distances – Figure 3.34 in Book 3 shows that the linearity of Hubble's law breaks down at large distances. The trouble with this is that the deviation from a straight line in Figure 1.6 appears noticeable only over large time intervals. That entails us measuring the distances of extremely distant objects, and this is very difficult. As we saw earlier, the distances are hard to determine with precision. None of the current observational measurements of this sort is sensitive enough to provide a definitive test of cosmological models.

Another test is based on the density of matter in the Universe. If we believe that all the matter in the Universe came into being at the Big Bang, then because the size of the Universe increases with time, the density must be falling. Therefore density is clearly a function of cosmic time. How do we measure the density of the Universe? We touched on this in Book 3, Chapter 4. Perhaps we should look for galaxies or clusters of galaxies and see how their number density in space changes with cosmic time. Alas, our telescopes do not allow us to penetrate sufficiently far back in time to use this as a precise test. As we probe further outwards, many galaxies become too faint to be seen.

The strongest evidence for density evolution comes from radio source counts. These counts were started in the 1960s. They firmly suggest that the Universe was denser in the past than at the current epoch, but again the data do not allow the precise form of the $R(t)$ versus t plot to be derived.

So, it all begins to look rather hopeless – except for one remarkable fact: the Einstein–de Sitter model of Figure 1.6 makes a firm prediction about the density of the Universe as it should be *today*. Forget about what it might have been in the past; just measure its present value!

Strictly speaking, to see how this conclusion arises, we ought to turn to the equations of general relativity theory. These, however, are far too difficult for this Course. There is, however, another plausible argument we can advance that is based purely on classical Newtonian gravity theory. This much simplified derivation just happens to come up with precisely the same answer. We present it below.

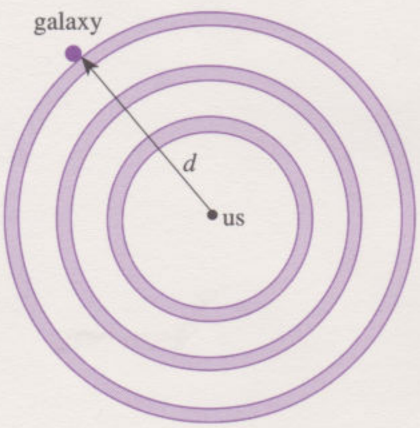

Figure 1.7 Spherical shells of matter centred on us. The gravitational force acting on the galaxy at distance, d, is the same as if all the matter in the shells was concentrated at the central point.

Let us assume the Universe is homogeneous and isotropic with a density given by ρ. Imagine a galaxy of mass m located at a distance, d, from us. Due to the expansion of the Universe, it is receding from us with speed, v. What are the gravitational forces acting on it tending to slow it down?

We regard ourselves as surrounded by shells of matter as shown in Figure 1.7 – the shells having their centre on our own position. Newtonian mechanics states that for a uniform spherical shell, the gravitational force it exerts on a test particle lying outside is that which would apply if all the matter of the shell were concentrated at the central point. If, however, the test particle lies within the region enclosed by the inner surfaces of the shell, the net gravitational force on the particle due to the shell is zero – that is to say, the different components of the force, arising from the different segments of the shell, exactly cancel. Thus, the gravitational force acting on the galaxy at distance, d, from us would be the sum total of the effects of all the shells, centred on ourselves, and having radii less than d.

The mass, M, contained in these shells is

$$M = \tfrac{4}{3}\pi d^3 \rho \tag{1.12}$$

Therefore the magnitude of the gravitational force on the galaxy due to this mass is:

$$F_g = GMm/d^2 \tag{1.13}$$

where G is the gravitational constant. The gravitational energy of the galaxy is

$$E_g = -GMm/d \tag{1.14}$$

and its kinetic energy is

$$E_k = \tfrac{1}{2}mv^2 \tag{1.15}$$

Thus its total energy E is

$$E = \tfrac{1}{2}mv^2 - GMm/d \tag{1.16}$$

As the Universe expands, d increases and v decreases. The density of matter decreases, but not the total amount of matter in the shells between us and the galaxy in question; in other words, M remains the same. This is because the Universe expands uniformly. In the Einstein–de Sitter model, in the very long-term future when d becomes infinite, the recessional speed, v, approaches zero.

□ If the Universe conformed to this model, what would be the total energy of the galaxy?

■ Both terms in Equation 1.16 would approach zero, so we are led to conclude that the total energy of the galaxy would also be zero.

Because the total energy is conserved, we may furthermore conclude that if the total energy of the galaxy will be zero in the far distant future, it must also be zero now and at every other instant. So we may set E to zero in Equation 1.16:

$$0 = \tfrac{1}{2}mv^2 - GMm/d$$

We can now substitute for M from Equation 1.12 and rearrange to obtain:

$$\rho = 3v^2/(8\pi G d^2) \tag{1.17}$$

But from Equation 1.4, v/d (in the Newtonian view) can be written as H_0, so

$$\rho = 3{H_0}^2/(8\pi G) \tag{1.18}$$

This same final equation also appears in the general relativistic treatment! Thus, we see that if the measured current values of the average density of matter, and of Hubble's constant satisfy Equation 1.18, the Einstein–de Sitter model would be the one to describe the Universe. If, on the other hand, ρ is either smaller or larger than the value required by Equation 1.18, then some other curve of $R(t)$ versus t would apply. The density given by Equation 1.18 is known as the **critical density**, and is usually denoted by ρ_c. In both Newtonian and general relativistic theories the critical density is the density of a universe with a scale factor tending towards a constant value as cosmic time tends to infinity.

So, everything depends on the measurement of the present-day value of ρ. We do not have to bother with what it might have been in the past. What could be simpler!

Unfortunately, as we shall see in Chapter 2, it is far from simple. A straightforward estimate based on the matter that is visible yields a value

considerably less than that given by Equation 1.18. And yet it is believed that the Universe *is* very close to critical density. (As you might have guessed, we would hardly have spent so long on one particular model that has the critical density if this had not been the case!) The reason is that there is evidence (as you saw in Book 3) of additional matter in the Universe apart from that which is visible. The intriguing significance of the dark matter, however, must be a story for later in this Block.

Summary of Section 1.4 and SAQs

1 At some point in the past, all the matter we see around us was concentrated in a region of zero volume and infinite density which then expanded. This is called the Big Bang.

2 The scale factor, $R(t)$, is the function of time by which all distances scale in a universe satisfying the cosmological principle. $R(t)$ tends to zero as t tends to zero.

3 For a freely expanding universe, $R(t)$ plotted against t is a straight line ($R(t)/R(t_0) = t/t_0$). In a universe with self-gravity, the expansion slows down and $R(t)$ versus t deviates from a straight line.

4 The deceleration parameter, q, is a measure of the rate of change of the gradient of the $R(t)$ versus t curve with time. $q = 0$ for a freely expanding universe and $q = \frac{1}{2}$ for an Einstein–de Sitter universe.

5 In the Einstein–de Sitter model of the Universe ($R(t)/R(t_0) = (t/t_0)^{2/3}$), the kinetic energy of the Big Bang 'balances' the potential energy of the galaxies. $R(t)$ tends to a constant as t tends to infinity.

6 An estimate of the age of the Universe, t_0, can be obtained by extrapolating backwards in time to the point on the $R(t)$ versus t curve where $R(t) = 0$.

$t_0 = 1/H_0$	freely expanding universe
$t_0 = 2/(3H_0)$	Einstein–de Sitter universe

(H_0 is the value of Hubble's constant *today*, at $t = t_0$.)

7 The critical density, $\rho_c = 3H_0^2/(8\pi G)$, is the present-day density of the Universe predicted by the Einstein–de Sitter model. Measurements of present-day density provide important tests of cosmological models.

SAQ 1.3 (Objective 1.4) If the possible values for H_0 lie in the range 50–100 km s^{-1} Mpc^{-1}, calculate the range of possible values for the age of the Universe, expressed in years, if it is undergoing expansion as an Einstein–de Sitter type of universe. (Do not forget to convert Mpc to km!)

SAQ 1.4 (Objective 1.3) For a universe given by Figure 1.6, the value of Hubble's constant does not remain constant with cosmic time. Why then call it a 'constant'?

SAQ 1.5 (Objective 1.1) Say why Equation 1.17 can also be used to calculate the escape speed of an object ejected from the Earth.

SAQ 1.6 (Objective 1.3) Calculate the values of the critical density for (a) $H_0 = 50$ km s^{-1} Mpc^{-1} and (b) $H_0 = 100$ km s^{-1} Mpc^{-1}. What would these translate to in terms of atoms of hydrogen per m^3, given that the mass of a hydrogen atom is 1.67×10^{-27} kg?

1.5 The resolution of Olbers' paradox

We are now in a position to resolve the puzzle of Olbers' paradox posed in Section 1.2. Remember the assumptions leading up to Olbers' paradox: an infinite and static universe uniformly populated by stars, a universe that is infinitely old. But now we see that the Universe is *not* infinitely old, nor is it static. Both of these factors contribute to the sky being dark at night.

Firstly, we know the Universe is not static but is expanding, therefore the light from the more distant objects is redshifted to a lower energy band.

☐ How does this affect Olbers' paradox?

■ As stated in the paradox, any line extending out into space will sooner or later meet up with the surface of a star. The light coming to us from that star was expected to be of comparable brightness to that of the Sun. (The total light received from this distant star will, of course, be less. This is because the disc of that star as seen from the Earth subtends a much, much smaller angle than the disc of the Sun. But within that small disc, the *brightness* is assumed to be comparable to that of the Sun.) However, we now know that the light from that distant star is redshifted. The photon energy is thus less, and so the light from a typical star is consequently *not* as bright.

You will recall from Book 1, Box 1.1, that $\varepsilon = hf$, where ε is the photon energy, f is the frequency of the light and h is Planck's constant.

Secondly, we see that the Universe is not infinitely old. It had a beginning, and so stars (and galaxies) have existed for only a finite time. Therefore, the maximum distance that light could have travelled is that which could have been covered in around 15 billion years. This means that to date we could have received light from only those stars, and any sources predating the formation of stars, contained within a sphere of radius 15 billion light years centred on the Earth. The surface of this sphere encloses what we call the **observable Universe**. Light emitted by any source lying outside the sphere has not been able to reach us yet, even if it started out immediately after the Big Bang. The surface of the observable Universe is depicted in Figure 1.8. As time passes, the surface expands farther outward at the speed of light; for each extra year added to the age of the Universe we receive light from a further light year away. In the observable Universe the galaxies and stars are not receding faster than the speed of light. This means that as the surface of the observable Universe expands it encompasses more and more galaxies, which now for the first time, become observable. But always the radius of the observable Universe is finite, and hence the number of stars in the observable Universe is finite (even though the number of stars in the *whole* Universe might be infinite).

☐ How does *this* factor contribute to a resolution of the paradox?

■ Although it might remain true that any line extended out into space intersects with a star *somewhere* in the Universe, it does not follow that the star must lie within the *observable* Universe. Because of the finite nature of the *observable* Universe, a line might encounter the boundary surface of our imaginary sphere of radius 15 billion light years before striking a star. Hence *no* light will have come to us yet from that particular direction – contrary to what was asserted in the paradox.

The finite age of the Universe is a more important factor than its expansion

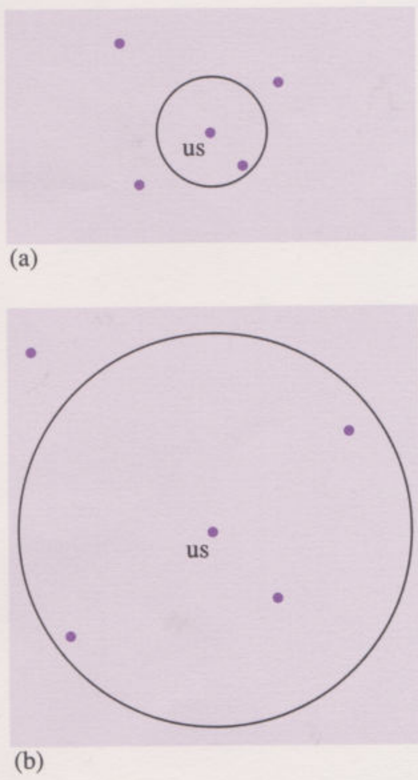

Figure 1.8 The observable Universe depicted (a) at cosmic time, t, and (b) at some later time, $(t + \Delta t)$. The galaxies shown are receding from us, but not faster than the speed of light, c. The observable Universe, therefore, encompasses more and more galaxies as time goes by.

In addition to Olbers' paradox, there was another conundrum we posed at the beginning: Why doesn't all the matter of the Universe pile up together in a heap because of its self-gravity? Again the Big Bang model of the Universe comes up with an explanation. Because the matter was given kinetic energy in the initial explosion at $t = 0$, the fabric of the cosmos expands with time; the galaxies are moving *away* from each other. In the Einstein–de Sitter model, the gravitational energy is precisely balanced by the kinetic energy but the galaxies never become static, only approaching this state, ever more slowly at greater cosmic times. You might care to speculate on what would happen if the Einstein–de Sitter curve of Figure 1.6 did *not* represent the Universe.

□ Suppose the density of matter in the Universe were *less* than the critical value in Equation 1.18 for the Einstein–de Sitter model. What would you expect to happen to the Universe eventually?

■ If the density is less than critical, the Universe does not become static even at infinite time.

Summary of Section 1.5

1 Olbers' paradox is resolved in the Big Bang model through two contributing factors:

(i) the redshift reduces the energy received from distant stars;

(ii) the finite age of the Universe means that the *observable* Universe is finite in size – not infinite as was assumed.

2 The Big Bang model of the Universe accounts for why all the matter of the Universe has not collapsed together.

1.6 The microwave background radiation

1.6.1 The Big Bang fireball

So, the expansion of the Universe is the first piece of evidence that the Universe began with a Big Bang. The motion of the galaxies is a direct consequence of that violent, cataclysmic event. We turn now to a completely independent observation that tells us that the Universe not only began with a Big Bang but with a *hot* Big Bang.

The discovery of the microwave background radiation in 1965 was one of the most fundamental and important events in astronomy. It ranks with the discovery by Hubble of the expanding Universe. Its story is a classic tale of the reward that comes from a full understanding of one's experimental apparatus. Two groups of radio astronomers had detected the radiation *previous* to the official discovery, but its significance had eluded them.

The discoverers, the American astronomer Arnold Penzias (1933–) and his student Robert Wilson (1936–), were employed by the Bell Telephone Laboratories in New Jersey. As part of their work on high-frequency radio receivers, they were using a special radio telescope in early experiments on

satellite communications (Plate 3.38). They were also using the telescope to do some radio astronomy and to map the emission from our Galaxy – at a wavelength of 7.35 cm. In testing its sensitivity, they found an unexpected signal. It seemed to be emanating from space. The radio waves did not vary in time over a six-month period, neither did they vary across the sky. Painstaking investigation ruled out the obvious sources of possible local emission. The signal definitely came from outer space – and most likely from beyond the Galaxy.

By coincidence, no more than a few hundred miles away, a second group of physicists had been arguing that the Big Bang must have been very hot and so should have been accompanied by a fireball – much like the fireball that occurs when a nuclear bomb is detonated. Remnants of this radiation ought still to be around today, filling the whole of the Universe. They calculated the strength of the radiation and commenced to build an experiment to search for it. However, before it could start operating, they realized that this could be the radiation that had been detected by Penzias and Wilson. They had been scooped!

Although the Bell Laboratories' astronomers were not searching for the **cosmic microwave background radiation**, as it came to be known, their ability to rule out precisely all possible local sources of radio emissions received by their telescope enabled them to be absolutely certain that they were detecting a uniform radiation from space, rather than radiation coming from specific stars or galaxies. For this historic discovery, Penzias and Wilson were awarded the Nobel Prize.

But what, you may ask, is the connection between the weak waves of *radio* frequency and the blinding light expected of the fireball?

Theorists calculate that for a Big Bang cosmological model, the early Universe should have been very hot, say 10^{10} K at a cosmic time of around 1 second. Radiation would have been in *thermal equilibrium* with all the matter in the Universe. Therefore the radiation would have had a black-body spectrum appropriate to that temperature (see Book 1, Subsection 1.3.2). All this happened a long time ago and we are still receiving the radiation. The radiation we see today has travelled for a long time: it originated near the spherical surface that marks the current boundary of the observable Universe (Figure 1.8). For reasons that will become apparent in Chapter 2, what we see now is a redshifted version of the original spectrum; every wavelength is increased by a common factor. An important characteristic of black-body spectra is that if you were to take a typical spectrum and increase all wavelengths by a common factor, the result is another black-body spectrum with the peak now at a longer wavelength. As we saw in Figure 1.14 of Book 1, the significance of having the peak of the distribution at a longer wavelength is that the spectrum now corresponds to a lower temperature. This figure is reproduced here as Figure 1.9.

We have touched on the idea of thermal equilibrium before. Here the important point is that the radiation and matter have the same temperature. This is because the matter and radiation interact strongly, i.e. they are *coupled*.

Therefore, in effect, what one sees today is a very much cooled down version of the original fireball. Theorists can calculate the extent of the cooling. Given the present baryonic density and age of the Universe, and assuming that no interactions have subsequently distorted the thermal equilibrium nature of the radiation, the radiation would have 'decoupled' from matter when the Universe was about 300 000 years old, and the temperature of the radiation at the current epoch should now be somewhat less than 5 K. Thus this cosmic background radiation should be easy to identify, having the typical shape expected of a black-body spectrum at that temperature. It should shine brightest in the very short wavelength microwave region, peaking at about 1 mm.

Remember that baryonic matter is 'ordinary' matter, consisting of protons, neutrons and some rare particles. (Electrons are not baryons, but the electron mass is nearly 2 000 times less than that of a proton, and there are equal numbers of protons and electrons in the Universe.)

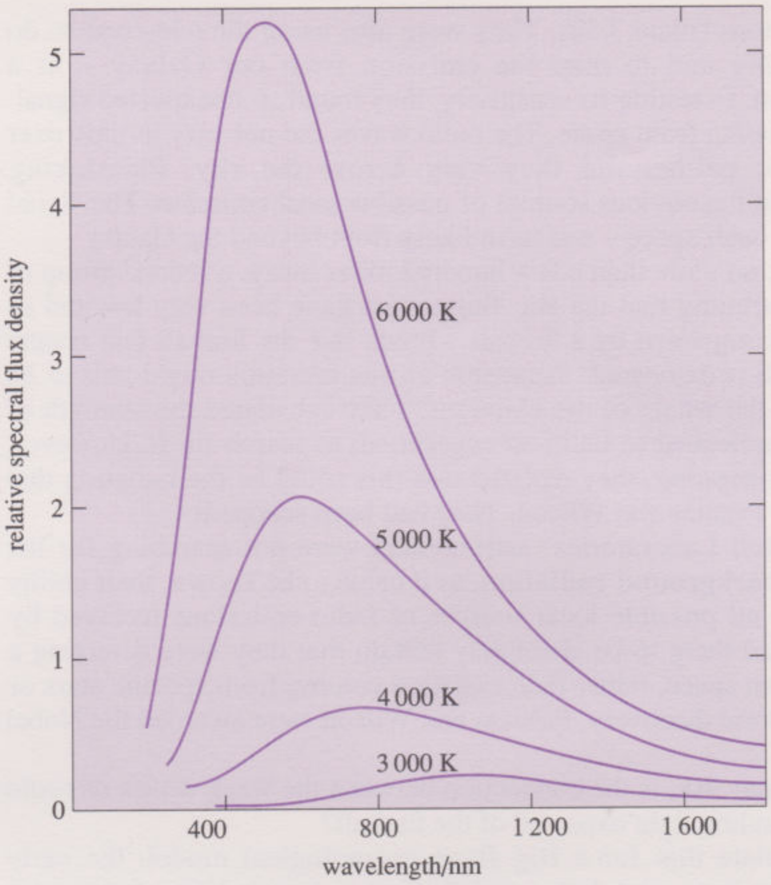

Figure 1.9 A series of curves of black-body radiation for various temperatures.

The measurements of Penzias and Wilson did not by themselves demonstrate conclusively that the radiation they had detected was the cosmic background radiation. After all, they had only detected radiation of a single wavelength. Nevertheless, it was isotropic and unvarying and a black-body curve of about 3 K could pass through their measurement. The search was soon on by many groups to measure this radiation at a variety of wavelengths. The results showed that on the longer-wavelength side of the spectrum, the experimental curve of spectral flux density versus wavelength fitted very closely that expected of a black body of around 3 K. The short-wavelength part of the spectrum on the other side of the peak is difficult to measure. This is due to atmospheric effects. (In Figure 1.25 of Book 1 the effectiveness of the Earth's atmosphere as a barrier to incoming radiation was shown.) A number of balloon-borne and rocket experiments produced conflicting results. In 1989 a satellite was launched that was dedicated to the study of the background radiation (Plate 3.39). That satellite, the Cosmic Background Explorer (COBE), away from the influence of the Earth's atmosphere, could detect the whole spectrum. This was crucial to show that the spectrum did indeed follow a black-body form – the results showed excellent agreement, as can be seen from Figure 1.10. The accuracy of the thermal spectrum is evidence for its origin at very early times.

The spectrum is an excellent fit to a black-body curve of 2.73 K – so good that there appears little reason to seek an alternative explanation. We conclude from this evidence that the Universe not only started with a Big Bang but that it must have been much hotter at earlier epochs; it was a *hot* Big Bang.

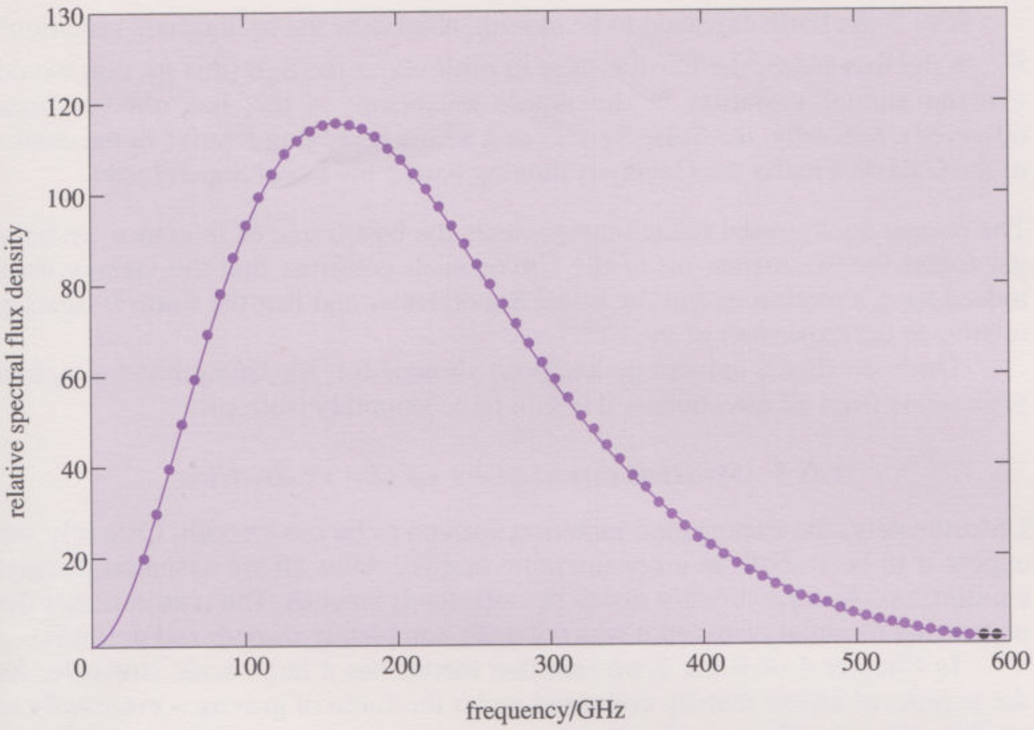

Figure 1.10 The spectrum of the cosmic microwave background radiation obtained by the COBE experiment in 1989 – the experimental data are the points, and the black-body spectrum at 2.73 K is the smooth curve. (Note that the *wavelength* at the peak here is 1.8 mm, *not* the 1.1 mm value on a graph versus wavelength. The reason is a bit subtle, and need not concern us, but arises from the fact that here the spectral flux density is per unit *frequency* interval, whereas in a graph versus wavelength it is per unit *wavelength* interval.)

1.6.2 Dipole anisotropy of the microwave radiation

So much for the confirmatory evidence of the Big Bang. What else might we learn from the microwave radiation?

□ Suppose an observer were to move relative to this background radiation. What effect might this have on the appearance of the radiation coming to the observer from different directions?

■ The radiation coming from a direction straight ahead, would be Doppler-shifted to shorter wavelengths – just as we would expect were we to approach any source of electromagnetic radiation. Similarly, the radiation coming from the opposite direction, directly behind the observer, would be Doppler-shifted to longer wavelengths. At all other angles the shifts would be intermediate between these two extreme values.

Starting in the 1970s, sensitive measurements were made using high-flying aircraft carrying millimetre-wave radiometers. They showed that the background radiation field in one direction was very slightly more intense than it was in the opposite direction. This effect is shown in Plate 3.40 in *Images of the Cosmos* taken from the latest data from the COBE satellite. This dipole anisotropy (so called because it differs in different directions) arises from the motion of the Earth with respect to the cosmic background radiation, i.e. relative to whatever spherical shell originally emitted the microwave radiation that the Earth is now receiving. The dipole proves that the radiation does not arise locally in sources that share our cosmic motion.

☐ Why is the Earth expected to be moving relative to the background radiation?

■ In the first place, the Earth moves in orbit about the Sun (this motion would give an annual variation in the dipole anisotropy – this has not yet been observed). Secondly, the Solar System as a whole is moving relative to the centre of the Galaxy. Finally the Galaxy is moving within the Local Supercluster.

The cosmic background radiation represents the best frame of reference we have yet found for the expansion of the Universe. It confirms that the Galaxy does indeed have a motion within the Local Supercluster and that the Earth is moving relative to the expansion of the Universe.

Once the dipole anisotropy has been allowed for, the background radiation is the same from all directions – it is said to be smoothly isotropic.

1.6.3 Spatial variations of the radiation

Unfortunately, the background radiation seemed to be *too* smooth. Certainly we expect it to be smooth to a considerable degree. After all we assumed thermal equilibrium. And yet it ought not to be *completely* smooth. The reason is that the matter that originally emitted it was not itself completely smooth and uniform.

In Chapter 4 of Book 3 we saw that matter has a large-scale structure. As the regions of higher density collapsed under the force of gravity – *eventually* to produce the superclusters of galaxies we see today – they would have heated up more than other regions of lower density, thus causing slight increases in temperature locally. You might expect these regions, therefore, to show up as 'hot-spots'. However, there is something of a complication at work here. Light loses energy when it has to drag itself away from a gravitational centre. So the light leaving these denser regions would have its wavelengths redshifted, i.e. this gravitational effect would tend to make the region look *cooler* than it really is. At the time of writing, it is not at all clear which of the two effects would win out. Would the denser regions appear hotter or cooler than normal – or could there be a mixture of the two, with different 'winners' in different directions?

Whichever way it goes, however, what we *can* say is that the radiation ought to show temperature variations of some sort over similar angular scales to those subtended by the superclusters today. In addition there ought to be variations on a smaller scale to correspond to the formation of individual galaxies within the supercluster. These temperature fluctuations ought still to be in evidence in the microwave radiation we observe today.

For a long time no such spatial fluctuations in temperature could be found. The situation got worrying. Was this throwing doubt on the interpretation that the radiation and matter had a common source in the Big Bang?

Then, in 1992, amid a blaze of publicity, it was announced that COBE had managed to measure spatial fluctuations in the cosmic background radiation. It was at an extremely low level – around 1 part in 100 000. The angular scale of the fluctuations was bigger (by a factor of about ten) than that subtended by superclusters. (This was because COBE measured intensity differences only between large regions subtending an angular scale of 10 degrees on the sky.) Nevertheless, the fluctuations were there. Plate 3.41 in *Images of the Cosmos* shows a picture of the intensity variations across the sky. These intensity variations are interpreted as temperature variations in the background radiation. In actual fact, this picture has to be interpreted with some care. It is not at all clear what the fluctuations in the cosmic background are. Instrumental 'noise' is expected to be responsible for producing quite a number of the observed features. (This 'noise' is somewhat similar to the snowy effect one sometimes gets on TV pictures.) As and when the experiment is repeated, what we expect to find is that certain of the features on this picture will be reproduced; others will not. The former can then be interpreted as the genuine signal; the rest as spurious 'noise'.

Thus, there is no incompatibility between the density distribution of matter and the temperature distribution of the microwave background radiation. Both show fluctuations at least on a scale greater than that of superclusters. It remains to be seen whether future measurements with finer angular resolution will eventually reveal fluctuations on the smaller scale corresponding to superclusters and clusters of galaxies.

Summary of Section 1.6 and SAQ

1 In a *hot* Big Bang the radiation in the early Universe would have been in thermal equilibrium with all the matter until the Universe was about 300 000 years old, and would have had a blackbody spectrum appropriate to the temperature at the time.

2 A redshifted version of this spectrum is another black-body spectrum corresponding to a lower temperature. At the present epoch the background spectrum, corresponding to a temperature less than 5 K, is predicted to peak at a wavelength of about 1 mm.

3 In 1965 Penzias and Wilson detected a microwave signal emanating from space which did not vary across the sky. This is the cosmic microwave background radiation, described in point 2.

4 The COBE telescope has detected the whole spectrum of the cosmic background radiation. The spectrum is a very accurate fit to a black body at a temperature of 2.73 K.

5 Dipole anisotropy in the background radiation reveals the motion of the Earth relative to the expansion of the Universe.

6 Slight spatial variations in the cosmic background radiation have been detected by COBE. These could correspond to spatial variations in matter, the denser parts of which later became superclusters.

SAQ 1.7 (Objective 1.5) Corresponding to each possible temperature of a black body there is a unique shape for the spectrum. Why then was it not sufficient to measure the spectral flux density of the background radiation at just one wavelength to see which black-body curve passed through it?

1.7 The hot Big Bang and nuclear abundances

There is a third piece of evidence for a hot Big Bang to add to the Hubble expansion and the cosmic background radiation. The Big Bang hypothesis provides a natural explanation of the abundances of the elements in the Universe. In this context, we are led to a convincing insight into the early history of the Universe. In particular we can be confident of extrapolating back in time to a point exceedingly close to the Big Bang; to a point when the Universe was only a few seconds old.

1.7.1 Conditions during the first 100 seconds

The evidence of the cosmic microwave background allows us to extrapolate back in time to when the Universe was 300 000 years old. It shows that the Universe was extremely hot at the early phases. In fact, there is a simple relation between

the temperature of the radiation field, T, and the scale factor $R(t)$, which we quote here without proof:

$$T \propto 1/R(t) \tag{1.19}$$

As we imagine going much further back in cosmic time to $t = 0$, $R(t)$ also tends to zero (Figure 1.6). Thus, the temperature tends to infinity.

If this extrapolation back in time is justified, it follows that at a time of around 1 second, the temperature was around 10^{10} K. The Universe was filled with matter and radiation colliding so frequently that, despite the rapid expansion, they were in thermal equilibrium. This being so, the behaviour of the Universe was dictated only by the prevailing temperature and the rules of thermodynamics. This is a key point; the nature of the Universe at about 1 second did not depend at all on what went before. We can therefore consider the events at this and subsequent times, even if we do not know in detail what happened before.

A positron, e^+, is an elementary particle with electron mass and positive charge equal to that of the electron. An antineutrino, $\overline{\nu}$, is an elementary particle with similar properties to those of the neutrino.

The matter at that time was composed of neutrons, protons, electrons and neutrinos, as well as positrons (antielectrons) and antineutrinos. For any particular elementary particle of mass m there was an epoch at which the thermal kinetic energy, of order kT (*Preparatory science*, Subsection 2.4.1), was equal to the rest energy, mc^2, of the particle (Book 1, Subsection 3.3.3). Because of thermal equilibrium, the photon energy would also have been of order kT. Up to this epoch therefore, a collision between two photons could result in the creation of a particle–antiparticle pair. (On collision, particle–antiparticle pairs can annihilate each other and convert into radiation.) For example, while temperatures were above 10^{10} K, electron–positron pairs could be created from the radiation field:

$$\gamma + \gamma \rightarrow e^+ + e^- \tag{1.20}$$

The symbol γ signifies a photon. In this case the high-energy photons must together have had an energy equal to at least twice the rest energy of an electron (0.511 MeV), a condition satisfied at temperatures above about 10^{10} K.

Remember that the electronvolt (eV) is a unit of energy: $1\,\text{eV} = 1.60 \times 10^{-19}$ J.

The small number of neutrons, n, and protons, p, present at this epoch interact with the positrons, e^+, electrons, e^-, neutrinos, ν, and antineutrinos, $\overline{\nu}$. The most common reactions are

$$n + e^+ \rightarrow p + \overline{\nu} \tag{1.21a}$$

$$p + e^- \rightarrow n + \nu \tag{1.21b}$$

Each reaction goes both ways and a balance between neutrons and protons is maintained. There is a small mass difference between the neutron and the proton – the neutron is the more massive by nearly three times the mass of the electron. The associated difference in rest energy corresponds to a temperature of 1.5×10^{10} K. As the neutron is more massive, reaction 1.21b needs *more* energy. This means that as temperatures drop below 10^{10} K, this reaction is inhibited, and this results in there being more protons and fewer neutrons. The ratio of the number of neutrons, n, to the number of protons, p, in the Universe, as determined by the reactions 1.21a and 1.21b, depends on the temperature and at equilibrium is given by

$$n/p = 10^{-(6.5 \times 10^9\,\text{K})/T} \tag{1.22}$$

which we quote here without proof.

ITQ 1.4 Confirm that the mass difference between the neutron and the proton corresponds to a temperature of 1.5×10^{10} K. (The mass of the neutron is $1.674\,82 \times 10^{-27}$ kg, whereas that of the proton is $1.672\,52 \times 10^{-27}$ kg. Boltzmann's constant, $k = 1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$.)

As the Universe expands after $t = 1$ second, the temperature drops below 10^{10} K, and the photons of the radiation field can no longer create electron–positron pairs. The electrons and positrons, however, continue to annihilate each other (the reverse of reaction 1.20). In addition, the energy of the neutrinos and antineutrinos drops and this makes them very unlikely to react with the neutrons and protons; they decouple from the matter, and are no longer in thermal equilibrium with it.

Free neutrons are unstable particles – they decay with a half-life of around 1 000 seconds into a proton and an electron and an antineutrino:

$$\mathrm{n} \rightarrow \mathrm{e}^- + \mathrm{p} + \overline{\nu} \tag{1.23}$$

The interaction is very unlikely to go the other way because this involves the collision of three particles simultaneously.

1.7.2 Nuclei building

So, we are faced with a situation in which the neutrons are decaying irreversibly into protons, and only one thing can save them: they must get absorbed into a nucleus. Only when bound inside a nucleus can a neutron be stable. Whereas at previous times, the violence of the collisions taking place immediately broke up any nuclei that had momentarily formed by fusion, now the conditions are calming down somewhat. As the temperature drops to 10^9 K, at $t = 100$ s, it becomes possible for neutrons and protons to fuse to form nuclei that are *not* disrupted. In this way, the neutrons and protons start to build up nuclei of deuterium (the isotope of hydrogen consisting of a neutron and proton, $^2_1$H, or D), and subsequently ~~lithium ($^7_3$Li) and ultimately~~ helium ($^4_2$He). × subsequently lithium ($^7_3$Li)

Note that, at these times, $T \propto \frac{1}{\sqrt{t}}$.

Thus, in essence, we find that up to about 100 seconds after the Big Bang, deuterium, and hence the larger nuclei, could not form because they kept on being broken up again. After a few thousand seconds, the free neutrons had decayed and were no longer available for deuterium production. *Therefore we are left with a very short time in which deuterium could be produced.*

As the temperature dropped, the ratio of neutrons to protons maintained by the reactions between neutrons and positrons and between protons and electrons (reactions 1.21a and 1.21b) had also dropped (Equation 1.22).

Reaction 1.21b declined with falling temperature, and reaction 1.21a declined as the supply of neutrons and positrons dried up. By the time all the various reactions stopped, matter was about 14% neutrons and 86% protons – the neutrons all incorporated into nuclei; the protons partly in the form of hydrogen, the rest incorporated into more complex nuclei along with the neutrons.

ITQ 1.5 Calculate the ratio of neutrons to protons given by Equation 1.22 for a Universe in thermal equilibrium at a temperature of 10^{10} K.

During the nuclei-building phase, once deuterium formed, most of it went on to produce $^4_2$He. Almost all the neutrons in the Universe in fact ended up in $^4_2$He. Nuclei heavier than $^4_2$He were not formed in appreciable numbers because no stable nuclei exist with 5 or 8 nuclear particles (for example the fusion product of a collision between $^4_2$He and a proton or between two $^4_2$He). There are the same number of protons in $^4_2$He as neutrons, so if *all* neutrons were incorporated into $^4_2$He, the mass fraction of $^4_2$He would be $(2 \times 14\%) = 28\%$. However, a small

fraction of neutrons will be bound in deuterium and other nuclei, resulting in the predicted mass fraction of ${}^4_2$He being somewhat lower – more like 25%. Unlike the case of ${}^4_2$He, the proportions of these other nuclei depend critically on the rate with which the particles were colliding during the nuclei-building phase. This, in turn, depends on the *density* of *baryonic* matter at that time. And the density of baryonic matter at that time can be determined from any assumed density of baryonic matter today and scaling appropriately to allow for the change in size of the Universe since then. Figure 1.11 shows how the proportions of the most abundant nuclei a few thousand seconds after the Big Bang are predicted to vary under different assumptions as to what the density of matter is today.

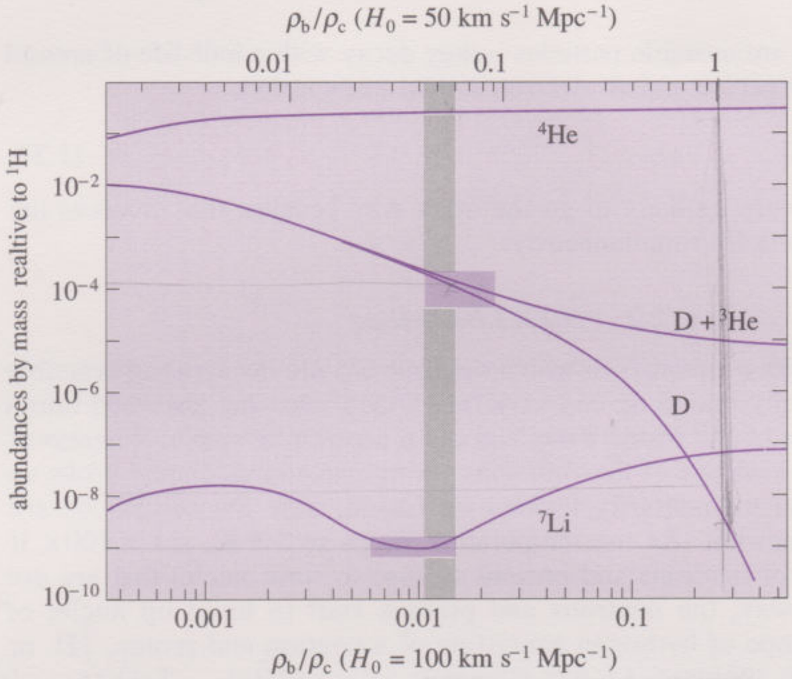

Figure 1.11 The abundances (by mass relative to hydrogen) of D, D + ^{3}He, ^{7}Li, and ^{4}He according to different assumptions regarding the density of baryonic matter today. ρ_b/ρ_c is the ratio of the present baryonic density of matter to the critical density. (The calculation assumes that the present temperature of the cosmic background radiation is 2.75 K. The upper scale for ρ_b/ρ_c assumes that H_0 = 50 km s^{-1} Mpc^{-1} and the lower scale that H_0 = 100 km s^{-1} Mpc^{-1}.) The observed range of abundances of D, D + ^{3}He, and ^{7}Li are shown. The vertical band indicates the range of densities consistent with all of these observations.

1.7.3 Further evidence for the Big Bang

But why, you might be asking, have we delved into this complex particle physics? It is because we believe that when we look out at the Universe *today* and sample the elemental abundances, the bulk of all the helium we find is due to that produced in the first few minutes of the Big Bang – nucleosynthesis in stars has still had no more than a rather modest effect, because most of the helium is locked up deep inside them. We can therefore ask what abundance of helium we should expect on the basis of Big Bang models, and compare it with this measured value. As we see from Figure 1.11, the theoretical proportion of helium does not depend sensitively on the baryonic density and hence which Big Bang model we adopt. Over a wide range of possible density values, the abundance of helium is expected to be approximately 25% by mass. And what is the value experimentally found? 25%! What an extraordinary way of confirming the Big Bang theory!

So, the observed helium abundance is exactly what we expect if the Universe began with a *hot* Big Bang, almost independent of the details of the baryonic density of the Universe. We can, therefore, be confident that we are indeed dealing with a *hot* Big Bang.

At the other extreme, the *deuterium* abundance is *very sensitive* to the baryonic density (Figure 1.11). This can also be very useful to us. On the assumption that we are indeed dealing with a Big Bang, we can use the measured abundance of deuterium to estimate the baryonic density – in other words, we turn the argument round. Deuterium is not produced inside stellar cores, so nearly all the deuterium we observe today must have come from the Big Bang. Deuterium can, however, be destroyed in stars, so today's observed abundance must be a lower limit on the value appropriate to the primordial abundance. The best place to measure primordial abundances is in the very cold reaches of interstellar space, where we hope the primordial mix has not been too contaminated by material thrown out from stars. Measurements of the abundance by mass of deuterium nuclei (that is, the ratio of the mass of deuterium nuclei to the mass of hydrogen nuclei) yield a value between 4×10^{-5} and 2×10^{-4}. The observed abundances of ^{7}Li and (D + ^{3}He) further limit the range of present-day baryonic density. Figure 1.11 indicates that the range of baryonic density consistent with all these observations is *very* small. The observed deuterium to hydrogen ratio is a lower limit, so it could be larger, and therefore the present-day baryonic density could be smaller. From Figure 1.11 we see that this points to a present-day baryonic density of about 0.05 ρ_c (upper scale), or 0.013 ρ_c (lower scale), where ρ_c is the critical density. (From Equation 1.18, this corresponds to a present-day baryonic density of about 2×10^{-28} kg m $^{-3}$.)

So, we see that the measured primordial helium and deuterium abundances give us a strong tool to constrain our cosmological models. They not only provide confirmation that the Universe did indeed begin with a *hot* Big Bang, but through allowing us to arrive at an estimate of the present baryonic density of matter, it has something relevant to add to our discussion in the next chapter about the ultimate fate of the Universe – something heavily dependent on the density of matter.

Summary of Section 1.7 and SAQ

1 The temperature, T, of the radiation field in the Universe, is proportional to $1/R(t)$, where $R(t)$ is the scale factor.

2 At t = 1 second, $T = 10^{10}$ K and matter and radiation were in thermal equilibrium. The conditions of the Universe at t = 1 second did not depend on what went before.

3 At t = 1 second the following reactions and their reverse reactions were important:

$\gamma + \gamma \rightarrow e^+ + e^-$

$n + e^+ \rightarrow p + \overline{\nu}$

$p + e^- \rightarrow n + \nu$

The ratio of neutrons to protons was maintained largely by these reactions.

4 By t = 100 seconds, the temperature had dropped to about $T = 10^9$ K. Electrons and positrons continued to annihilate each other. The neutrinos and antineutrinos decoupled from the matter. Free neutrons decayed with a half-life of 1 000 seconds.

5 When the temperature dropped to about 10^9 K nuclei began to form. Protons and neutrons were bound into the nuclei of deuterium. The n/p ratio was frozen.

6 Almost all the neutrons ended up in $^4_2$He. The fraction by mass of helium is expected to be 25%, which is in agreement with observation.

7 The abundances of D, ^{3}H and ^{7}Li in the Universe are very sensitive to the baryonic density of the Universe during the nuclei-building phase. The abundances of these nuclides therefore provide a strong tool with which to estimate the density of the Universe to test the cosmological models.

SAQ 1.8 (Objective 1.6) From our discussion, what can you conclude about the origin of the *heavy* elements (heavier than ^{4}He) we see in the cosmos today?

1.8 An overview of the Big Bang

We have now presented the three reasons why we believe there was a hot Big Bang. Let us now summarize how we think the Big Bang developed.

To do this we need to choose a starting time. We cannot begin our description at $t = 0$ because, according to the ideas we have presented, all the matter of the Universe would have infinite density and infinite temperature. This is not a situation our current formulation of physics can handle! We need to start at a time where things have had a chance to get themselves sorted out to a reasonable extent. Fortunately, as we have already pointed out, the results of the nucleosynthesis early in the development of the Universe do not depend on $t = 0$ because thermal equilibrium prevailed for a time. We begin at $t = 1$ second.

Now that in itself might strike you as pretty ambitious. A Universe that has been around for 15 billion years, and we are saying what was happening when it was just one second old!? Strange to say, on the basis of the physics to which you have already been introduced, we can be reasonably confident in our description of the Universe from that time to the present day. Indeed, on the basis of modern physics, we shall in the next chapter be extending the scenario right back to $t = 10^{-43}$ s and forward into the indefinite future!

As already mentioned, the temperature at $t = 1$ s is thought to have been 10^{10} K. That is several orders of magnitude higher than what is today found in the centre of main sequence stars. At that time, not only were there no molecules or atoms, there were not even any nuclei – only neutrons, protons, electrons and neutrinos, (and their antiparticles), together making up a plasma. As we saw in the previous section, deuterium and helium formation did not begin to take place until the Universe was a few hundred seconds old. The Universe continued its expansion. The density decreased steadily and the radiation and matter cooled. Yet they remained in thermal equilibrium due to the electromagnetic interactions between the photons and the protons (and other nuclei) and electrons. The Universe still resembled a thick soup of interacting constituents.

Things changed dramatically as the temperature slowly fell to around 10^4 K at $t = 10\,000$ *years*. Normally we think of the process we are about to describe in reverse. If we take a box of hydrogen gas and heat it up, then at a certain temperature, around 10^4 K, the atoms will start to become ionized as the electrons detach themselves from the protons and we end up with an ionized plasma. This is just what we have in the early Universe, an ionized plasma of protons and electrons, plus some helium nuclei and a trace of other nuclei. But as the temperature cooled to this critical temperature, the electrons and nuclei for the first time began to form *atoms*. Once this started to happen, it proceeded very quickly. Suddenly, the Universe changed from an ionized plasma to a neutral gas

composed mainly of hydrogen and helium atoms. This transition had a drastic effect on the photons of the radiation field. Previously they interacted strongly with the charged particles, changing direction and energy with every interaction. While this was the case, the Universe was opaque (just another way of saying that photons travelled only a small distance before they interacted with matter). Once the Universe became atomic in character, however, the radiation field no longer interacted with the material and the Universe became transparent to photons. This happened at $t = 300\,000$ years.

The photons are then said to have *decoupled* from the matter. They now travel through matter without suffering any significant interactions. These photons are the photons we now detect as the cosmic background radiation. They provide direct information from the epoch in the Universe when the radiation and matter decoupled. This is referred to as the *decoupling epoch*. Now we see why the smoothness of the cosmic background is so important. This tells us precisely about the smoothness of the matter at the time the hydrogen atoms formed. Superclusters of galaxies must have formed much later, from the small inhomogeneities present at the decoupling epoch.

The decoupling epoch is also (somewhat misleadingly) called the recombination epoch.

We are now entering what might be called the epoch of observation – after all, we can see and measure the cosmic background radiation. The gaseous material collected together to form superclusters of galaxies, and within each galaxy the first generation of stars ignited. In the interior of these stars, primordial hydrogen and helium were fused to produce heavier elements. These were then expelled by supernovae into the interstellar medium. There they became recycled into later generation stars, such as the Sun, and into planets, such as the Earth. And that – very briefly – is how, 15 000 million years after it all began, we come to be here today, piecing the story together.

Objectives for Chapter 1

After studying Chapter 1 (and any associated audio, video or TV material), you should be able to:

1.1 Give brief definitions of the terms, concepts and principles listed at the end of the Objectives.

1.2 Explain the nature of Olbers' paradox and how this is resolved.

1.3 Describe the expansion of the Universe, particularly in terms of Hubble's law.

1.4 Describe how the age of the Universe can be determined, and state the current best estimates.

1.5 Describe the microwave background radiation, and explain why this is evidence for a hot Big Bang.

1.6 Discuss how the primordial abundances of the elements can be accounted for in terms of nuclear processes that took place in a hot Big Bang.

1.7 Provide a summary of the Big Bang scenario, giving an indication of the time-scales involved in the various processes.

List of scientific terms, concepts and principles used in Chapter 1

Term	Page	Term	Page
Big Bang	12	Einstein–de Sitter model	14
cosmic microwave background radiation	21	homogeneous (universe)	10
cosmic time	12	Hubble's law	8
cosmological principle	10	isotropic (universe)	10
cosmology	6	observable Universe	19
critical density	17	Olbers' paradox	7
deceleration parameter, q	14	scale factor, $R(t)$	12

Chapter 2
The origin and evolution of the Universe

Prepared for the Course Team by Russell Stannard

2.1 Introduction

In this chapter we seek a deeper understanding of the Big Bang and of what the future might hold for the Universe. In so doing we draw heavily on the ideas contained in Einstein's general theory of relativity. These ideas form the bedrock of modern cosmology.

It is customary for treatments of general relativity to be highly mathematical. But take heart; that will *not* be the case here. Instead, we shall be content merely to present you with the fundamental thinking behind the theory, making use of everyday analogies wherever possible, rather than the customary mathematics.

General relativity is a fascinating subject; it completely alters the way we view the Universe. Instead of thinking exclusively of objects, such as galaxies, exerting gravitational forces upon each other across an otherwise empty, passive arena of space, we have to regard space itself as an active participant in the drama. And it is not just our conventional notion of space that is challenged by relativity; the nature of time has also to be re-evaluated. What emerges is not space and time as separate entities, but an extraordinary amalgam of the two, so that there can be no space without time, and no time without space.

It is only through such revolutionary thought that certain cosmological questions can be answered. Indeed, certain questions that come readily to mind, like 'What happened *before* the Big Bang?' are exposed by relativity as possibly having no answer at all. Why? Because the question is meaningless!

One final note before we start: No, there is no truth in the rumour that only Einstein and one other genius ever understood relativity. Relativity is for everyone to understand and enjoy.

2.2 The nature of the expansion

The impression we have probably given you in Chapter 1 is that the Big Bang went as follows:

All the matter in the Universe suddenly appeared at a certain point in space, went off bang, and ever since has been spreading out, progressively filling up the rest of space.

A very natural assumption – but one that is quite wrong. The Big Bang is not like any other explosion that has ever taken place. Sure enough, all the matter of the *observable* Universe started off crowded together at a point. But not only all of its matter; all of its space was contained at that point too.

What then happened was that the *space itself* began to expand. As it did so, it carried the matter out with it. The process is still carrying on today. Why are the galaxies receding from each other? Because the space between them is expanding, and in so doing, it draws them further and further apart.

Crazy! How can *nothing* move anything? The answer is that, to a physicist, space is *not* nothing. When talking of 'empty space' all we mean is that there aren't any lumps of matter such as stars and planets around; we do not mean to imply that there is absolutely nothing there. To a physicist, empty space is to be regarded as a uniform continuum. The word 'uniform' is the key to why we are normally unaware of the stuff that makes up space. After all, how do we know that something (like this book you are reading) exists? Because you can point to it; you can say 'It is there – and not over there, or there'. It is trickier to convince yourself that air exists because it is all around you. But you can do it – by pursing your lips and blowing, for example. The fluttering of a piece of paper held up in front of your mouth is evidence enough that the air is there. But note what you

are doing. You are changing the density of the air so that there is a greater concentration of it in one region than another. In other words, you have destroyed its uniformity. To understand what physicists mean by space, all you have to do is take the analogy of undetectable uniform air a step further. Suppose there was a kind of stuff that permeated everything. And we do mean *everything*: all of space between objects, and all of space *inside* objects – even the space inside the atoms (the space separating the electrons from each other and from the central nucleus). And suppose this stuff were absolutely uniform everywhere. Would you be able to tell that it was there? No. For instance, as you walked through it, unlike the air, it would remain uniform in density – there would be no build-up of pressure in front tending to resist your motion; it would simply slip through you unnoticed.

But if this background continuum cannot be detected, what is the point of suggesting that it is there? The interesting thing is that it *can* be detected; under certain circumstances, its uniformity *can* be disturbed. Subnuclear particles, for example, sometimes produce antiparticles in high-energy collisions. Antiparticles can be regarded as holes that have been knocked out of the background continuum. Not only that, but one might expect that an electrically charged particle, such as an electron, would attract particles of the continuum that had the opposite charge to itself, and repel those that had a like charge. In other words, each charged particle would disturb the otherwise uniform distribution of charge in the continuum. This effect has also been observed and measured.

Thus one finds that, whether physicists are dealing with the motion of the Universe's largest constituents, or the quantum behaviour of its smallest, it becomes much more natural to think of space as a (normally) completely uniform distribution of stuff, rather than as nothing. In either field of study, it is not a passive stage on which particles perform like actors, but is itself an active participant.

That being the case we return once more to our consideration of the Big Bang:

Imagine, for a moment, blowing up a balloon and that – for some curious reason – before doing so, one gets a tube of superglue and sticks an array of 5 p coins to the surface of the balloon. Assuming this fiddly operation was successfully managed without glueing the fingers together, one starts to blow up the balloon (see Plate 3.35). The 5 p coins separate from each other as the rubber sheet between them expands. The coins are being borne along by the expanding rubber; it is not a case of the coins moving *through* the rubber, or sliding over its surface.

One has here a good analogy for the expansion of the Universe. Just as the coins are borne along by the rubber, so the galaxies are being borne along by the expansion of space; they are not moving *through* space.

□ Suppose a fly were to alight upon one of the 5 p coins as the balloon is expanding. It sees all the other coins receding from its own. From this observation, it concludes that it must have alighted, by chance, on the coin at the centre of the expansion. How could a fellow fly convince the first of its mistake?

■ It could point out that if the first fly were to join him on another coin, the view looks the same. It doesn't matter which coin one is on, all the others appear to be receding away from oneself.

The same is true for observers in the galaxies. All the other galaxies are receding from our galaxy. But, as we noted in Chapter 1 in the discussion around Figure 1.4, this does not mean that we are at the centre of the Universe. The same claim could be made by anyone living in any other galaxy.

One point that can be a little confusing is that not *all* distances expand. This is the case when objects exert physical forces on each other so as to form a bound

system. So, for example, the Earth is bound to the Sun by gravitational attraction. As such, they are to be regarded as a single object, so the distance between them does *not* expand along with the general expansion of space. The same is true of the stars that make up a galaxy; they too are to be treated as making up a single object – the galaxy. So, it is only the distances *between* bound objects that expand, not the objects themselves. In this, respect, the analogy of the coins on the rubber balloon is very close; the coins do not increase in size, only the distances between them.

ITQ 2.1 A fly resting on one of the coins sees that another coin 50 mm away is receding from it at a speed of $10\,\mathrm{mm\,s^{-1}}$. How fast will be the speed of recession for coins placed at (a) 100 mm and (b) 300 mm away from the fly?

20
60

As you will have seen from the answer to ITQ 2.1, the speed of recession of the coins on the balloon is proportional to their distance from the observer. But that was just what we found to be the case for the galaxies.

2.2.1 *The origin of the redshift*

In Chapter 1, this proportional relationship between speed and distance arose out of Hubble's law, through the Hubble redshift being interpreted as a consequence of a recessional speed.

When we spoke of this interpretation, we were regarding the Universe from a Newtonian point of view. The redshift was treated as a normal type of Doppler shift. But, as we warned at the time, all was not well with that explanation. What was wrong with it? Now that we have switched models to one based on general relativity, we must re-examine the origin of the redshift.

As an example of a normal Doppler shift, take the case of the pitch of a police van's siren. As it passes your position and retreats into the distance, the frequency of the sound goes down. If the van were stationary, successive humps and successive troughs in the wave train would be separated by the wavelength, λ, where $\lambda = c_s/f$, c_s being the speed of sound in air, and f the frequency of the siren. But with the van on the move away from you, it is passing through the air – the medium carrying the sound waves. During the time between the emission of one hump and the next ($1/f$), the van has moved through this medium a distance v/f, where v is the speed of the van. So, as can be seen from Figure 2.1, the successive humps are now separated by a total distance of $\lambda + v/f$. This longer wavelength, in turn, corresponds to a lower frequency.

The cosmological redshift has an entirely different origin. The galaxy is *not* moving through any 'medium' for carrying the light wave. An observer in that galaxy, moving with the source, will see the light being emitted with its *normal* wavelength. That being so, how is it, when we later receive that light, it has an increased wavelength?

□ A distant galaxy has emitted light of its normal wavelength towards us. What do you think happens to this light *during* its journey?

■ The journey takes time. During that time, space expands – including the space between successive humps in the wave train of light.

So, although an observer on a distant galaxy can attest that the wavelength of light had its usual value at emission, this wavelength progressively lengthened on its way towards us. This in turn meant the frequency went down. The further the light travelled, the more pronounced the lengthening of the wavelength, i.e. the further away the galaxy is from us, the lower will be the frequency of the light we receive from it. Such then is the origin of the cosmological redshift.

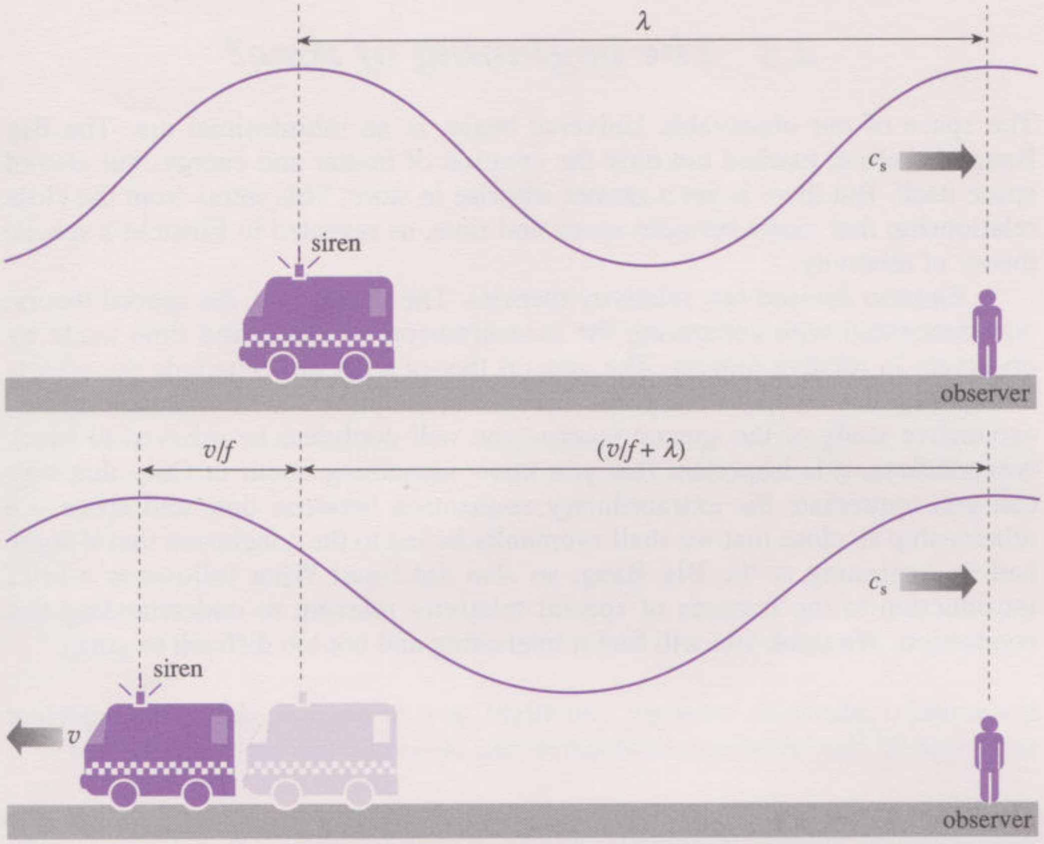

Figure 2.1 The wavelength of the sound emitted by a siren moving away from an observer is longer than it was when stationary because the successive humps and troughs in the wave train are emitted at different locations.

Identical considerations apply to the cosmological microwave background radiation. This radiation started out its journey 15 billion years ago with a short-wavelength spectrum characteristic of a *hot* Big Bang. But in all the time it has taken to reach us, the wavelengths have been increased by the expansion of the intervening space through which it has passed.

This would be a good point at which to view video sequence 13, The expanding Universe. *Remember to read the associated notes first.*

Summary of Section 2.2 and SAQs

1 The expansion of the Universe is due to the expansion of space itself.

2 It is this expansion of space, rather than a Doppler shift, that explains the cosmological redshift.

SAQ 2.1 (Objective 2.2) Suppose someone looks around at space today, and asks you at which point the Big Bang occurred. What help would you be able to offer? (You might like to make use of the rubber balloon analogy.)

SAQ 2.2 (Objective 2.2) A distant galaxy is moving relative to the centre of the local cluster of galaxies to which it belongs. This motion is directed towards us here on Earth. To an observer on Earth, what effect, if any, would this motion have on the wavelength of the light (a) as it is emitted and (b) as it is received by us?

2.3 The beginning of time?

The space of our observable Universe began as an infinitesimal dot. The Big Bang, therefore, marked not only the creation of matter and energy, but also of space itself. But there is yet a greater surprise in store. This stems from the close relationship that exists between space and time, as revealed in Einstein's special theory of relativity.

Einstein devised *two* relativity theories. The earlier one, the special theory, was concerned with comparing the measurements of space and time made by observers in relative motion. The general theory went on to include the effects that gravity has on space and time. This is not the place for us to launch into an exhaustive study of the special theory (you will doubtless be relieved to hear). Nevertheless, it is important that you know something about it. Only that way can you appreciate the extraordinary connection between time and space – a relationship so close that we shall eventually be led to the conclusion that if space had its beginning at the Big Bang, so also did time! What follows is a brief introduction to the features of special relativity relevant to understanding this connection. We think you will find it interesting and not too difficult to grasp.

If you find it otherwise, however, you might, as a last resort, skip to the marginal note on p. 42 that alerts you as to where you should be sure to rejoin the text.

We begin by imagining an astronaut about to undertake a space journey to a distant planet. The first surprising consequence of relativity is that the time of the journey, as measured on a clock in the space capsule, would be less than that measured on an identical clock belonging to a mission controller at Houston. This is not because there is something wrong with one of the clocks. Both clocks are keeping perfect time. No, the reason for the disagreement is that *time for someone moving in the space capsule is not the same as time at Houston.* Only observers that are not moving relative to each other share the same time.

This strange state of affairs applies not only to exotic undertakings such as space journeys, but *whenever* two observers and their clocks are moving relative to each other. The fact that in normal everyday life we can get away with the notion that we all share a common time is due to the convenient fact that the differences in the two times becomes pronounced only at speeds close to that of light. This is just as well – otherwise we would have to reset our watches every time we hopped on a bus!

Convenient? Yes. But also misleading. Because we are normally unaware of the differences between our times, we grow up under the false impression that there is only one time – THE time! This is something that has to be unlearned.

This slowing down of time for the observer moving relative to us is no tentative hypothesis; it is an experimentally verified fact. And the measured slow-down rate, at any particular relative velocity, agrees exactly with that predicted by Einstein's theory for that velocity. Einstein's relationship linking the astronaut's time, t_a, to the controller's time, t_c, is

$$t_a = t_c(1 - v^2/c^2)^{\frac{1}{2}} \tag{2.1}$$

where v is their relative speed, and c is the speed of light. (This we state without proof.)

ITQ 2.2 What would be the ratio of t_a/t_c for $v = 0.9c$?

From ITQ 2.2, we discover that an astronaut travelling at 9/10th the speed of light would have his time slowed down to about half the rate of that of the mission controller.

What if the speed of the astronaut exceeds the speed of light? That would give a negative value for $(1 - v^2/c^2)$ in Equation 2.1, and the square root of a negative number is no real number! In fact this problem does not arise. A further consequence of relativity theory is that nothing can travel faster than light. The speed of any object, v, must always be less than c. The reason for this is that the faster an object goes the more energy it has, and therefore according to special relativity, the greater the amount of extra energy needed to give the object a given increase in speed, i.e. it becomes harder to accelerate.

One thing the astronaut and the mission controller agree about is their relative speed. The controller sees the astronaut disappearing off into the distance at the same speed as the astronaut sees the Earth receding from him. But doesn't that raise a problem? Won't the astronaut, with the 'slow' clock, be surprised by the short duration of the journey? Won't the astronaut be able to calculate that it is not possible to have travelled all the way to that distant planet in the time recorded on the space capsule clock? No. It turns out that our astronaut is perfectly happy to accomplish the journey in the observed time. Why? Because of a second consequence of the special theory:

□ Given that the astronaut and controller agree about the speed of the capsule, but they disagree over the time of the journey, what does that imply about the distance of the journey, as measured by the astronaut?

■ We know that *distance* = *speed* × *time*. If the two speeds are the same, but the astronaut's time is less than that of the controller, then the astronaut's estimated distance to the planet at the start of the journey must be smaller. Thus, for example, at 9/10th the speed of light, the astronaut's time is about half – but the planet is only half the distance away.

Different times, different distances. It all sounds so confusing. But that really is not the case; it doesn't have to be any more confusing than other situations where people make differing observations.

Suppose, for example, you were to hold up a stick in a room where a meeting of amateur photographers was taking place, and you asked everyone to photograph it from where they were sitting. If the stick were held horizontally, each photograph would look different.

□ Explain in your own words why this would be so.

■ What a photograph shows is merely a two-dimensional projection, at right angles to the line of sight. In reality, the object is three-dimensional (Figure 2.2).

The reason we do not regard the differing projections 'confusing' is that we have learned that these are but *appearances*; they do not show the true length of the stick. To obtain that, one has to recognize that the stick not only has an extension in the two-dimensional plane at right angles to the line of sight, but also a component of length in the third dimension lying *along* the line of sight. So, for example, the person seeing the stick almost end-on in Figure 2.2 loses a large component along the line of sight, leaving a comparatively small projected length at right angles. For someone viewing the stick from the side in Figure 2.2, the component along the line of sight is a small fraction of the large projected length at right angles.

The method of working out the true length is as follows. Each observer takes the line of sight to be his/her Z direction: Z_1 or Z_2. Two additional axes at right angles are chosen to be the X and Y directions, so giving $X_1\,Y_1\,Z_1$ or $X_2\,Y_2\,Z_2$ (see Figure 2.3). The true length of the stick, l, is then given in terms of the projections, x, y, and z along the three axes:

$$l^2 = x_1^2 + y_1^2 + z_1^2 = x_2^2 + y_2^2 + z_2^2 \qquad (2.2)$$

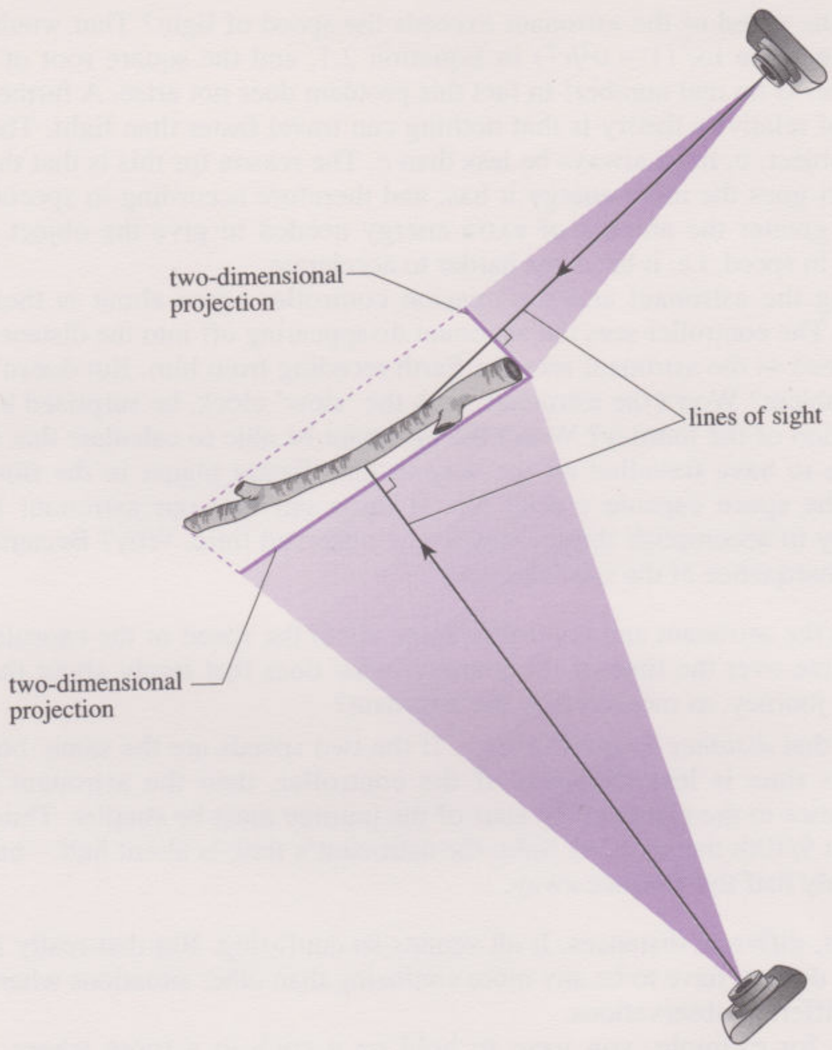

Figure 2.2 The line of sight to the stick varies according to the relative positions of the photographers, so the projections will look different. A photograph taken from the side will show a long length of stick; another taken end-on will show a short stubby length.

No two observers will have the same values of x, y, and z because they each have their own choice of X, Y, and Z axes, depending on their vantage point relative to the stick. But this does not matter; using their own values of x, y, and z, they arrive at identical values of l. It is the consistency of these values of l that confirms us in our belief that the stick is actually three-dimensional, even though superficially it might appear to be only two-dimensional.

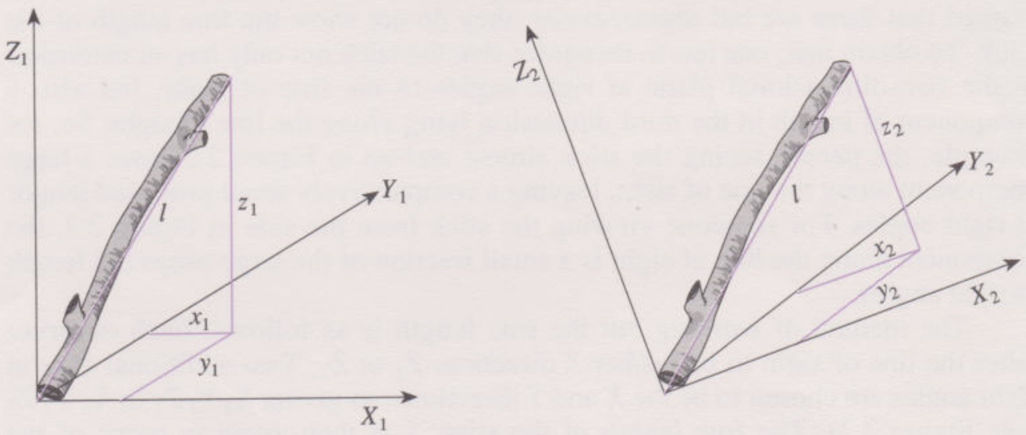

Figure 2.3 Regardless of which set of axes one adopts, $X_1\,Y_1\,Z_1$, or $X_2\,Y_2\,Z_2$, the length calculated for the stick comes out to be the same in both cases.

That being so, we return to the case of the astronaut and the mission controller. How are they to reconcile their differing journey distances and times? The answer from the special theory of relativity is to recognize them as simply *appearances*. According to the theory, space and time are not separate entities; they are merely *projections*. True reality is four-dimensional. There are three spatial dimensions and one temporal dimension. This is no playing around with words. What we mean is that the time dimension is seamlessly welded onto the space dimensions to make up a four-dimensional continuum called **spacetime**. There can be no space without time, and no time without space – a statement that will assume great significance when we shortly return to explore further the Big Bang.

What prompts us to make such a seemingly outrageous claim? The supposed four-dimensional object is the astronaut's space journey. We note that it is bounded by two events: the launching of the spacecraft from Earth at a given time, and its arrival at the planet at a different time. As far as the controller is concerned, the two events are separated by a certain distance and time. For the astronaut, the events occur at the same place – at the spacecraft – so are separated by no distance at all (motion being relative, it appears to him that he is sitting still as the Earth recedes from him and the planet comes towards him.). He regards the time difference between the two events to be about half that of the controller's (assuming as before that the relative speed is 9/10th that of light). Each observer now regards his measured distance and time as mere projections. Combining them to produce an estimate of the true separation in four dimensions, *they arrive at exactly the same value for the separation!* The formula they have to use is this:

$$s^2 = c^2t^2 - x^2 - y^2 - z^2 \qquad (2.3)$$

It is analogous to the one we had before – Equation 2.2. The quantity, s, is the separation, or 'distance', in four dimensions between the two events. The expression for s has four terms, whereas previously we needed only three. The first term is the new one and takes care of the component of separation along the time axis. The time difference between the two events, t, has to be multiplied by a speed so that we can express this component in metres like the three other components, instead of the usual seconds. The speed that needs to be chosen for the theory to work is c, the speed of light in a vacuum. The last three terms are simply what we had before – the square of the three-dimensional spatial distance, l^2, but this time with a negative sign in front. This negative sign accompanying the three spatial terms – compared to the positive sign of the temporal term – serves to underline the fact that, although relativity theory reveals similarities between space and time that we would otherwise not have suspected, there remains, nevertheless, something distinctive about them.

You may be wondering why we didn't keep the square of the spatial distance positive, as before, and simply subtract a fourth term, c^2t^2. The reason is that c^2t^2 is always larger than l^2 for any journey we might undertake. Writing it as we have done ensures that s^2 comes out positive, so it becomes a straightforward matter to take the square root to find the separation, s.

So, to sum up, what we are saying is that the individual values of l and t are mere appearances depending on the observer's vantage point. In the case of the photographers, the different vantage points arose from their sitting in different positions in the room; in the present case, it arises from the different relative *motions* of the two observers. Just as the photographers can change the appearance of the stick at will by changing their location, so astronauts can change their perception of a journey by changing their speed. But no matter what speed is adopted, when they come to evaluate s, they will always get the same answer for the distance of the journey in four dimensions – and this will agree

with the estimate obtained by the mission controller. It is the fact that there is always agreement over the value of s that leads us to accept that s describes the real state of things – not l or t.

Having said that, however, a word of reassurance is called for. If you are experiencing difficulty forming a mental image of what four dimensions looks like, don't worry – it cannot be done! Also you should not try contorting your thumb and three of your fingers to represent the four axes all mutually at right angles to each other – that also cannot be done. But that doesn't matter. Mental pictures and tangible models are helpful wherever possible – but not necessary. All the physicist *needs* is Equation 2.3.

This is where you should rejoin the text if you have skipped the discussion of special relativity.

And so it is we arrive – at long last – to the reason behind this rather lengthy excursion into special relativity.

☐ Can you guess what its relevance might be for our consideration of the Big Bang?

■ Earlier we explained how the Big Bang saw not only the creation of matter and energy, but also of space. But what we have now learned is that you cannot have space without time, or time without space. Therefore, if the Big Bang marked the beginning of space, *it also marked the beginning of time*.

This being so, we find ourselves quite unable to tackle questions such as: What happened before the Big Bang? This is not because we do not know the answer, nor that there are reasons that lead us to believe that we shall *never* know the answer. Rather, the point is that *the question itself* is meaningless. There was no 'before'.

An immediate consequence of this is that we have to be extremely careful when we approach questions that deal with what might have *caused* the Big Bang. Normally when something happens we regard it as an effect caused by something else. Cause leads to effect; that effect in turn becoming the cause of the next effect down the causal chain. Thus, for example, careless child knocks cup → cup falls → cup breaks → child gets told off → etc. But we are unable to regard the Big Bang as just another link in this chain – this link, like all others, having its own cause. Such a cause would have had to have occurred earlier than itself – but there was no 'earlier'.

Imagining oneself going back in time and finding the time axis coming to a halt at the Big Bang is interesting enough. But the British physicist Stephen Hawking (1942–), author of the best-selling popularization of the subject, *A Brief History of Time,* has proposed an even stranger scenario. According to his speculation (and that is really all it is at present), as one goes back in time towards the Big Bang, the character of time changes; it progressively becomes more like space. Earlier we noted that, although relativity showed up unsuspected similarities between time and space, time nevertheless retained a certain distinctiveness. What Hawking is suggesting is that under the extreme conditions of high temperature, colossal gravitational forces, and small distances (allowing quantum effects to dominate), the distinction might vanish as we penetrate deeper and deeper into the realm of the Big Bang. It could be that time in effect 'melts away'; one never reaches any definite beginning point.

Regardless of Hawking's speculation, the lack of any time before the Big Bang appears to do away with a God who lights the blue touch paper and retires. Succumbing to a common urge experienced by many popularizers of science to wax philosophical in the later chapters of their books, Hawking, in his best-seller, asks rhetorically, 'What place for a creator?'. Whilst it would not be appropriate in an Open University science course to launch into an extended section on theology, a word or two might be in order (especially as we are nearing the end of *our* account of cosmology!).

In the first place, no self-respecting theologian would hold to such a naive view of creation. In the theology of the major world religions the concept of the *Creator* is nearly always linked closely to that of the *Sustainer*. The notion of creation is one that applies to *all* points in time – not just to the first instant (if there was one). The emphasis is not on how everything got started, but *why is there something rather than nothing* – an important distinction. This latter question is one that appears to be as relevant now as ever.

A suggestion put forward by certain scientists is that the Universe spontaneously created itself. How? Through a quantum fluctuation. On the small scale – normally associated with subatomic physics – quantum physics reigns. Because of the famous Heisenberg uncertainty principle, we find that subatomic particles are not subject to the usual type of deterministic causal chain. Instead, one can do no better than predict the probabilities of *various* possible outcomes. The argument goes that because the observable Universe began small, everything happening would have been governed by quantum physics. For all we know, out of a state of nothing, there might have been a non-zero probability of a universe appearing! In other words, the Big Bang occurred spontaneously. It is an interesting, though non-proven suggestion.

But is it the solution to all our problems? No. It leaves the question of why the laws of quantum physics were in charge rather than some other conceivable form of physics. Where did the *laws* come from? Did it require Someone to choose?

Incidentally, before we leave the subject of the beginning of time, it is interesting to note that there is actually nothing new in the idea. Saint Augustine argued that, time being a feature of the world, God had to create it along with everything else. In other words there was no time before the world existed. It is a salutary thought for modern cosmologists that this Christian theologian got there some 1 500 years before they did!

Summary of Section 2.3 and SAQ

1 It would seem that the instant of the Big Bang marked not only the creation of matter/energy, but also of space.

2 The special theory of relativity reveals that space and time are inextricably linked together in a four-dimensional spacetime. Thus, if the instant of the Big Bang marked the creation of space, it also marked the creation of time. Accordingly, it becomes meaningless to speak of what happened before the Big Bang – there was no 'before'.

3 There still remains the question of why there is something rather than nothing. Even if the Universe created itself spontaneously through a quantum fluctuation of some kind, that would still leave open the question as to why the laws governing the process were those of quantum physics rather than of some other conceivable form of physics.

SAQ 2.3 (Objective 2.3) Using the ratio you obtained in ITQ 2.2 for t_a/t_c, together with Equation 2.3, check that the astronaut and controller with relative speed $0.9c$ do obtain the same value of s for the separation between the two events marking the beginning and end of the space journey.

2.4 The future of the Universe

So much for the beginning of the Universe; what of its end? What does the future hold?

The answer depends on the density of matter/energy in the Universe. In Chapter 1 we introduced the Einstein–de Sitter model. This was the one where the density of matter had a value such as to provide just sufficient mutual gravitational attraction to succeed in halting the expansion at an infinite time in the future. This value of density, you remember, is known as the 'critical density'. If the actual value is greater than this, then the expansion of the Universe will be halted after a finite time. We call this type of Universe a **closed universe**. If, on the other hand, the density is less than critical, the expansion will continue for ever. This we call an **open universe**. (The reason for the names 'closed' and 'open' will become clearer in Section 2.5.) Let us now examine each of these possible scenarios in turn.

2.4.1 Closed universe

As we said, this is a universe where the density is sufficiently great that the expansion one day comes to a halt. But that's not the end of it. Gravity is still operating. So what happens next is that all the matter in the Universe begins to be dragged back together. Slowly at first, but with ever increasing speed, the galaxies rush towards each other. Space is now contracting.

□ What would you expect to happen to the microwave background radiation?

■ Just as the expansion of space during the initial phase caused the wavelength of the radiation to stretch out, and hence the temperature of the radiation became reduced, so now the contraction of space causes the wavelength of the radiation to decrease, and the corresponding temperature to rise.

As the contraction continues, the sky becomes as bright as a star's surface. The Big Bang scenario now follows in reverse order, with the atoms breaking up under the continual impact of the radiation; then the nuclei getting smashed apart into their component nucleons. Finally, the nucleons disintegrate, releasing their constituents: point-like objects called quarks. Thus, we end up with a soup of quarks, electrons (and less familiar particles), which then squashes down into a **Big Crunch**.

What happens after that? Who knows? Physics as we know it at present gives up at this point, and we are left guessing. One possibility is that that would be the end of everything. Matter and energy, space and time – everything goes out of existence.

But there is an alternative. There might be a **Big Bounce** – with all the matter rushing out again. The cycle would then be all set to repeat itself. This is not to say that the events of one cycle are replicated again; we as individuals will not relive our lives. But there would be another period of expansion, galaxy formation, etc., followed by a final drawing together again. We would get a whole series of Big Bounces. What we have so far been calling the Big Bang would be merely the most recent of the Big Bounces.

□ What would be the significance of this for what we said earlier about the beginning of time?

■ If there have been previous Big Bounces, then clearly what we said earlier about the Big Bang marking the beginning of time must be revised; there would have had to have been a 'before' in order to accommodate the earlier cycles.

Thus, according to this picture, the Universe could have existed for ever. There need be no instant marking creation, because there need be no beginning to the sequence. (This, of course, still leaves us with the question as to why there is a universe at all, i.e. something rather than nothing.) There need also be no end to the sequence; the Universe could keep on bouncing for ever.

At present we have no way of deciding which of the two alternatives – annihilation or bounce – would win out. Some scientists might regard the eternal bouncing scenario as aesthetically the more pleasing, but that is not to say they have any idea of what kind of force would convert the attraction of gravity, which is the cause of the crunch, into the repulsion needed to produce the bounce.

2.4.2 Open universe

If the density is less than the critical value, then the Universe will continue to expand for ever. The future of the Universe will be something like this:

After an estimated 10^{12} years, all the hydrogen, helium, and other light elements in the interstellar medium get used up, so no more stars form. A point is reached where nothing is left except cooling white dwarfs, neutron stars, black holes, planets, rocks and meteoroids.

These will continue to orbit the centres of their respective galaxies until some such time as they might undergo collisions that upset their motion and lead to them being drawn to the centre of the galaxy. Even without such collisions, it is expected that the bodies will continually lose energy, resulting in a spiral path that gradually takes them ever closer to the centre. (This is because they emit a form of radiation associated with gravitational forces.)

□ What might happen to them when they reach the centre of the galaxy?

■ They could get swallowed up by a massive black hole that progressively eats up more and more of the galaxy.

□ What do you think the final mass of such a black hole might be?

■ Assuming that it swallows up all the stars of the visible galaxy, the black hole should reach a mass roughly equal to that of the once-visible matter in the galaxy: $10^{11}M_{\odot}$ for our galaxy (Book 3, Subsection 1.2.2). [Estimates indicate that this might occur in roughly 10^{27} years.]

The galaxies themselves lose energy (through the emission of gravitational radiation) as they move about each other in their galactic cluster. They therefore tend to be drawn to the centre of their galactic cluster where, in about 10^{31} years, they form a supergalactic black hole with a mass of order $10^{15}M_{\odot}$.

All this points to a cold, lifeless universe of black holes. Except that even that is not quite the end of the story. As has been pointed out by Stephen Hawking, an interesting consequence of quantum theory is that the matter/energy locked up in a black hole does not remain there for ever. It is beyond the scope of this Course to explain how this comes about. Suffice it to say that the extreme variations in gravity close to a black hole cause subatomic particles to form out of the space. If this happens just outside the hole, then there is a chance that some of these particles will be produced with just enough energy to drag themselves away from the vicinity of the hole. This is done at the expense of the mass/energy of the hole itself. Thus, in effect, the hole evaporates. The time of evaporation is governed by the mass of the hole. Black holes like that in Cygnus X-1 will evaporate in 10^{67} years; those at the centres of galaxies in 10^{97} years; and supergalactic black holes in 10^{106} years.

Thus, the final end of an open universe is that it consists of slow-moving subatomic particles progressively becoming ever more dilute as the expansion continues. This is the so-called **Heat Death** of the Universe.

Of course, the end might not arise exactly in accordance with the scenario outlined above. For one thing we have said nothing yet about **dark matter** – that extra matter required to help galaxies form (Book 3, Subsection 2.5.1). This is largely because (as we shall see later) we do not know for sure what kind of matter it is.

Then there might be new phenomena. For example, there is the possibility that protons might decay into positrons and neutrinos before they have a chance to get absorbed into black holes. If so, slow-moving baryonic matter would become 'hot' (i.e. fast-moving) smaller particles. The lifetime for such decays – if they occur – is expected to be in excess of 10^{30} years (anything less and we should have noticed it happening). For most purposes, this is so long that the effect of the decays can be ignored. But on the time-scales we try to deal with in projecting the future of the Universe, such a process would have a significant effect on the scenario.

All this having been said, however, the general nature of the end seems pretty clear.

2.4.3 Is the Universe open or closed?

So, which scenario is it to be? We have already seen (Subsection 1.4.5) that the most direct way of answering the question, by measurement of the deceleration parameter, q, yields inconclusive results; the uncertainties are simply too great to allow us to decide. The same applies to trying to measure the way the density has changed over time. Instead we must argue on the basis of whether the estimated density today is greater than, or less than the critical value. If we take an intermediate value of H_0 of $75\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, the critical value is about $10^{-26}\,\mathrm{kg\,m^{-3}}$, or the equivalent of 6 hydrogen atoms $\mathrm{m^{-3}}$.

If we add up how much matter we can actually *see*, mainly in the form of stars, we arrive at a density of about 1% the critical value – clearly a long way off.

□ Does this mean we can immediately conclude that the Universe is open?

■ Recall what we learned in Book 3, Subsection 2.4.2. From the rotation curves of spiral galaxies, and the motion of galaxies in clusters of galaxies, it is possible to infer that there is a lot more matter than can be seen in the form of luminous stars. We called it dark matter. Dark matter will contribute to the overall density and so has to be included in the inventory.

This additional dark matter could take the overall density to 20% critical, or possibly more.

Finally, there is the question of whether there is matter that goes beyond even the clusters of galaxies and influences the spatial distribution, and the motion, of galaxies on an even larger scale. As was discussed in Book 3, Chapter 4, the answer is known to be yes, and this will bring the density even closer to the critical value.

What is the nature of this dark matter? The halos of the galaxies are possibly made up of black holes and brown dwarfs, the latter being collections of matter that were not quite massive enough to raise the temperature sufficiently to ignite nuclear reactions – in other words, 'failed' stars (Book 1, Subsection 3.3.4). But we have to be careful. Such matter would be baryonic – made largely from neutrons and protons. There is a limit on how much of that type of matter we can have around before we start upsetting the primordial nuclear abundances. Looking back to Figure 1.11, you will note that the abundance of deuterium, ${}^{2}_{1}\mathrm{H}$,

depends very sensitively on the assumed baryonic density (in contrast to the abundance of helium which hardly depends on the density at all, and for that very reason provides excellent confirmation of the Big Bang theory). As we saw in Subsection 1.7.3, the abundance of ${}^{2}_{1}\mathrm{H}$ yields a present-day density of baryonic matter of less than about $2 \times 10^{-28}\,\mathrm{kg\,m^{-3}}$, i.e. *less* than a few per cent of the critical value. There is such confidence in the nuclear physics involved, that we must find some other type of dark matter to make up the bulk of the dark matter. What might that be?

There are, of course, large numbers of photons: 550 million of them in every $\mathrm{m^3}$. Their energy (indeed any energy) has a gravitational effect. But their energy density amounts to a gravitational effect only 10^{-3} that of the luminous matter.

A more promising candidate is the neutrino. It was originally thought to be massless. But, as noted in Book 3, Section 4.3, perhaps it is not quite so. We certainly know that the mass has to be small. An analysis of the time of flight of the neutrinos detected from the 1987 supernova, compared to how long it took the light of the supernova to reach us from that distant galaxy, shows that the neutrinos must have been travelling exceedingly close to the speed of light. This in turn puts limits on the mass of a neutrino. The observation of neutrinos from this supernova yields an upper limit to the mass of 2×10^{-5} that of the electron.

ITQ 2.3 If there were 10^8 neutrinos to every baryon, and they each had a mass of 2×10^{-5} that of the electron, what would they contribute to the density of matter in the Universe?

Estimates of the number of neutrinos compared to the number of baryons give values between 10^8 and 10^{10}. Thus, as we have seen from the answer to ITQ 2.3, it would be quite possible for a neutrino mass of order 10^{-5} that of the electron to provide enough dark matter (*hot* dark matter in this case) to reach the critical value. However, it does have to be made clear that at the present time there is no direct evidence that the neutrino has any mass.

Not only that, there is an additional problem. Recall from Book 3, Subsection 2.5.1, that dark matter seems to be necessary for galaxies to form from more dispersed matter. Whereas the fast-moving ('hot') neutrinos might well allow this on a large scale (the formation of galaxy clusters), they do not seem to have the ability to cause matter to collapse on the smaller, more local scale characteristic of individual galaxies – at least, not in the time that was available for this. Thus, it would appear that neutrinos cannot be the only type of dark matter.

An alternative is to postulate the existence of slow-moving 'exotic' types of matter – types of particle, left over from the extremely high-energy reactions of the Big Bang, that are as yet unknown to us. Various elementary particle theories do indeed require the existence of types of matter additional to that with which we are familiar. Suffice to say that such 'cold' dark matter would serve to ensure the collapse of matter on the scale of individual galaxies. However, computer simulations show that it would not be effective at forming clusters and superclusters of galaxies.

Thus, by way of concluding our discussion of possible forms of dark matter, we can state the following: the most likely outcome, to which we alluded in Chapter 4 of Book 3, is that there is a mixture of different types. There is more baryonic matter than that which is luminous (but less than a few per cent critical), plus some exotic matter to ensure galaxy formation, plus massive neutrinos to form the clusters. One preferred cocktail is 70% 'hot', to 30% 'cold' dark matter. This would work very nicely. Nevertheless, it must be stressed that all such talk

is for the present speculative. (The next, revised, version of this book might have to have a hefty substitution at this point!)

And as for our discussion of whether the Universe is open or closed, how are we to summarize that? We can be fairly sure that the density lies between about 10% of the critical value and about double the critical value. But this, of course, still leaves the options wide open. However, that is not the end of the matter – as you will see when we once again take up the subject in Section 2.6.

Summary of Section 2.4 and SAQ

1. The future of the Universe depends upon how much matter/energy it contains.
2. Above the critical density, the expansion will one day come to a halt; thereafter, the Universe will collapse together in a Big Crunch. That might be the end. Alternatively, there might be a Big Bounce, in which case, the Big Bang might merely have been the most recent Big Bounce of a perpetually oscillating Universe.
3. If the density is less than the critical value, the expansion will continue for ever and the Universe will suffer a Heat Death.
4. If it has *exactly* the critical value, the Universe will just halt its expansion in the infinite future. Again, during that span of time, it would have suffered a Heat Death.
5. Estimates of the density show that it has a value close to critical (within a factor of 10). One cannot be sure of its exact value because it is mainly in the form of dark matter. This could amount to 100 times the luminous matter contained in stars. Dark matter in the form of baryons cannot exceed a few per cent of the critical value; the rest could be made up of unknown types of particle left over from the Big Bang, and/or neutrinos having mass.

SAQ 2.4 (Objective 2.4) Suppose that the density in the Universe is greater than critical, such that at some point in the future all the galaxies come to a halt. What would happen to the radiation given out by the galaxies as they start to come together? Would the appearance of the galaxies, as seen from Earth, change and if so, would such changes occur at the same time, or not?

2.5 The geometry of spacetime

Earlier, when dealing with the expansion of the Universe, we spoke of space expanding and carrying the galaxies along with it. But then in the last section we reverted to the more familiar idea of the galaxies moving under the influence of their mutual gravitational forces. We ended up with a hybrid theory whereby the large-scale movement of the galaxies was governed by the properties of spacetime, and the more local motions by gravity forces. This is not necessary. Indeed, in the general theory of relativity we do away entirely with the notion of gravity as a force. Instead, we treat *all* such movements as being governed by the properties of space – or more exactly, the properties of spacetime.

The motivation for this is simple. According to Newtonian ideas, the 'natural' state of motion for any object that is moving is that it should continue to move in a straight line at a steady speed. Alternatively, if the object is initially at rest, it should remain at rest. This state of affairs is regarded as 'natural' in the sense that it does not seem to call for any kind of explanation. Only if we see

some departure from this behaviour – the moving object deviates from a straight line or changes its speed, or the object at rest suddenly starts to move off in a particular direction – only then do we feel compelled to seek a reason. And the reasons offered by Newtonian theory are in terms of forces acting on the bodies. The degree to which the object deviates from the 'natural' state of motion depends both on the strength of the force and on the mass of the object.

As far as the gravitational force is concerned, this way of looking at things immediately throws up a problem. Two objects of different mass, released from rest at the same time from the same height, will arrive at the ground at the same time (neglecting the effects of air resistance). This simultaneous arrival indicates that the acceleration of the two bodies was identical. But to accomplish this feat, gravity had to pull harder on one object compared to the other.

□ Object A has a mass of 3 kg; object B a mass of 12 kg. In order that their accelerations under gravity should be identical, how much harder must gravity pull on B compared to A?

■ Force = mass × acceleration. Thus, if the mass of B is 4 times that of A, but the acceleration is to be the same, the force on B must be 4 times that on A.

A variant on this is to consider a spacecraft in orbit around the Earth. With its engines off it follows an elliptical path. An astronaut steps outside the craft for a space walk, and floats alongside the craft in essentially the same orbit. But to keep the craft and the astronaut on the same path, gravity must pull on the craft with a greater force than that acting on the astronaut – again in the exact ratio of their masses. How does gravity manage this trick?

Einstein was greatly struck by this strange coincidence. It was this that led him to propose the general theory of relativity. According to Einstein, the 'natural' state of motion is not necessarily one of being at rest, or of moving in a straight line at a constant speed. For the two bodies released from the same height above the ground, the natural motion is *not* that they should stay put until acted upon by a force, but rather that they should accelerate towards the ground with the observed value of acceleration! And not just those two bodies; *any* body placed at that position would experience the same acceleration. Mass does not come into it; neither do gravitational forces. The bodies simply follow the natural motion associated with that region of space. The governing factor is not any force, but *the properties of spacetime itself*. The acceleration is a characteristic of that particular location. Space close to a body like the Earth, is not the same as space elsewhere – it is warped.

In a similar vein, we must think of the spacecraft and the astronaut also 'doing what comes naturally', rather than being constrained to follow some other path by the imposition of an external force. We must learn to think that for objects (*any* objects) travelling at that velocity, at that particular point in space, their natural motion is to follow the observed orbit – not to go flying off into space in a straight line.

A helpful analogy is to think of high-speed racing cars going round a racing track – a *banked* track. On such a track, the natural motion is not to go in a straight line. If the banking is ideal, you ought to be able to complete a circuit shouting 'Look! No hands!' There would be no need to apply sideways forces through operating the steering wheel; the design of the track itself would do the job for you.

So it is with the spacecraft and astronaut. In effect, they are going round their circuit shouting 'Look! No gravity!'. They are being kept on track by the design, or the properties, of the spacetime in that region.

The situation we have in mind is similar to the two-dimensional analogy shown in Figure 2.4. The ball-bearings cause dents in the otherwise flat surface of the rubber sheet. Another ball rolling past will be drawn towards the ball-

bearing; another might be following a closed path about it (neglecting the slowing-down effects of air resistance and friction). If we were to look directly down on these movements – so we were unable to see the distortion to the rubber surface – we might conclude that the curvature of the paths was due to some hypothetical horizontal force between the balls and the ball-bearings. But this would be mistaken; the balls are simply tracing out the natural paths characteristic of the distorted space over which they are rolling. The force on the ball is that exerted by the curved surface at the point of contact between the rubber and the ball; it is not an action-at-a-distance force exerted by the ball-bearing on the ball.

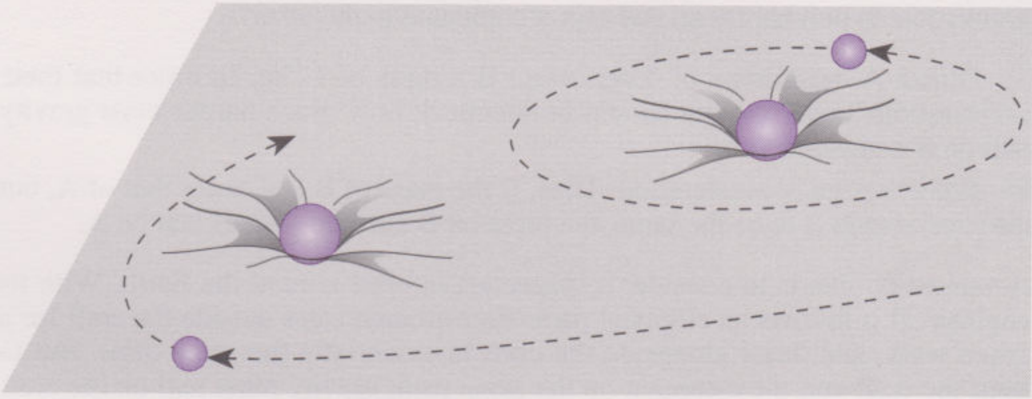

Figure 2.4 The paths of rolling balls being affected by the indentations made on a rubber sheet by heavy ball-bearings. (For clarity, the small indentations around the rolling balls are omitted.)

The ball-bearings, of course, represent bodies like the Earth and the Sun. They too distort the space in which they are situated in such a way that the motion of nearby bodies is affected. In this case, instead of a distorted two-dimensional surface, it is three-dimensional space that is affected.

Helpful though this rubber sheet model might be, the analogy must not be pressed too far. In truth, it is not just three-dimensional space that is affected by gravity, but four-dimensional spacetime. The presence of matter affects not only the properties of space but also of time. General relativity requires that time close to a body such as the Sun runs more slowly than time further away! The fact that a passing body thus spends 'longer' in the vicinity of the Sun than it would otherwise have done according to Newtonian ideas, adds further to the amount by which the passing body's trajectory is deflected. This is an effect that has been experimentally verified.

Thus, we are able to replace entirely the concept of a gravitational force by a distortion, or warping of spacetime. When we do this we find that general relativity comes up with *more accurate* results than the old theory.

For example, there is an aspect of the movement of the planet Mercury that cannot be explained in any other known way. Mercury is the planet that probes the region of space closest to the Sun, where the effects of spacetime distortion in the Solar System are the most pronounced. It follows an elliptical path. According to Newtonian ideas of gravity, the path of a small, isolated planet in orbit around a spherical Sun should exactly repeat itself. But that of Mercury does not; the perihelion (the point of closest approach to the Sun) progressively advances (see Figure 2.5). Mostly this can be explained as the influence of the other planets. But a residual amount (43 arcsec per century!) is left unaccounted for in terms of Newtonian gravity. Einstein, however, was able to show that just such an effect would be expected on the basis of his theory of relativity, and moreover the size of the effect expected agrees accurately with that which is observed.

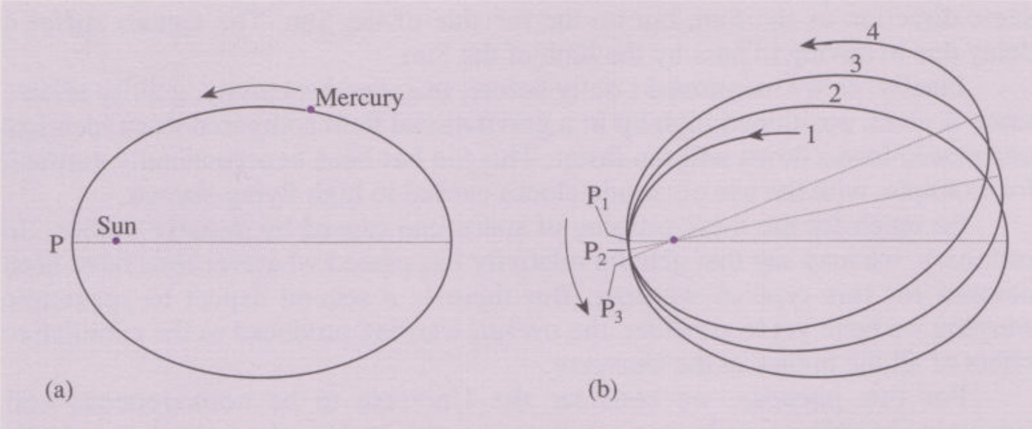

Figure 2.5 (a) The orbit of an isolated planet around the Sun, according to Newtonian mechanics; (b) The advance of the perihelion of the planet Mercury (exaggerated). This can be explained by Einstein's theory of relativity. (~~For clarity, the small indentations around the balls are omitted.~~)

Further confirmation of the correctness of general relativity comes from the behaviour of light. If spacetime is warped, then *everything* in this region should be affected by the distortion – including light. For example, light from a distant star, passing close to the Sun on its way to us, ought to follow a bent path. As the Sun passes through that region of the sky where the star is situated, the position of the star ought to appear to alter, depending on how much the light path is bent. Of course, it does not take much thought to realize that this is not likely to be an easy observation to make. Normally one sees stars at night-time – not when the Sun is about! What one had to do, in fact, was wait for a total eclipse. Then, for a brief spell one could see the stars by the light that passes close by the Sun. When this was done, what one found was that their apparent positions were indeed different from those they are observed to have at other times of the year (when the Sun is in some other part of the sky). The amount of the bending was found experimentally to agree with that predicted by the theory (see Figure 2.6).

Nowadays, one does not have to wait for a total eclipse. Instead, one makes use of radio sources and measures by how much the radio waves are bent by the Sun. This can be done whether or not there is an eclipse.

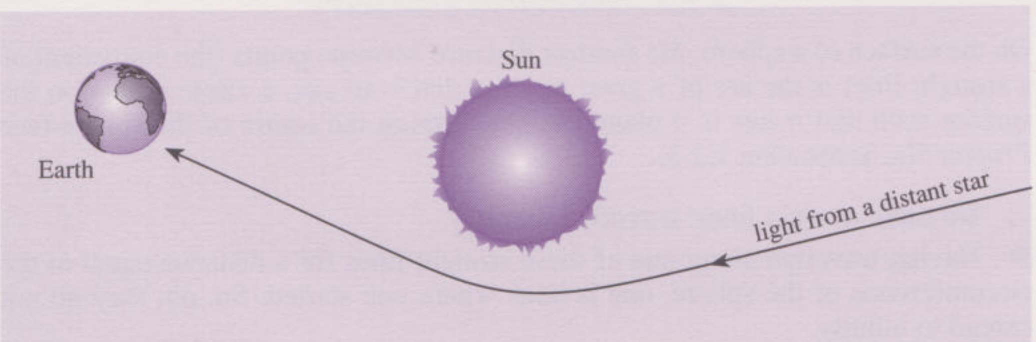

Figure 2.6 Light from a distant star passing close to the Sun follows the bent path shown.

The reason for using the Sun is, of course, the need for a source of strong gravity to produce a noticeable effect on the light or radio waves. Even so, the maximum bending is no more than 1.75 arcsec. More recent observations, however, have succeeded in measuring the even smaller light-bending caused by the planet Jupiter.

Yet another consequence of general relativity is that light passing through a region of space close to a source of gravity should be slowed down. This too has been verified by bouncing radar signals off planets when they lie in nearly the

same direction as the Sun, but on the far side of the Sun. The signals suffer a delay due to having to pass by the limb of the Sun.

Finally, as we mentioned briefly before, in general relativity, gravity affects time. A clock positioned high up in a gravitational field compared to an identical one placed lower down will run faster. This too has been experimentally verified, for example, with the use of atomic clocks carried in high flying aircraft.

So much for the local warping of spacetime caused by massive bodies. In summary, we may say that general relativity has passed whatever tests have been devised for this type of warping. But there is a second aspect to spacetime warping we have yet to consider: the *overall* warping produced as the cumulative effect of all the bodies in the Universe.

For this purpose, we consider the Universe to be homogeneous and isotropic. In other words, we are ignoring the local indentations caused by individual stars, or by galaxies, or even by clusters of galaxies. We imagine the matter smoothly distributed throughout space causing a constant curvature everywhere.

There are just three possibilities for the overall geometry of such a homogeneous spacetime. These depend on the mass density of the Universe (to which we shall return shortly). Their analogues for two-dimensional surfaces are shown in Figure 2.7.

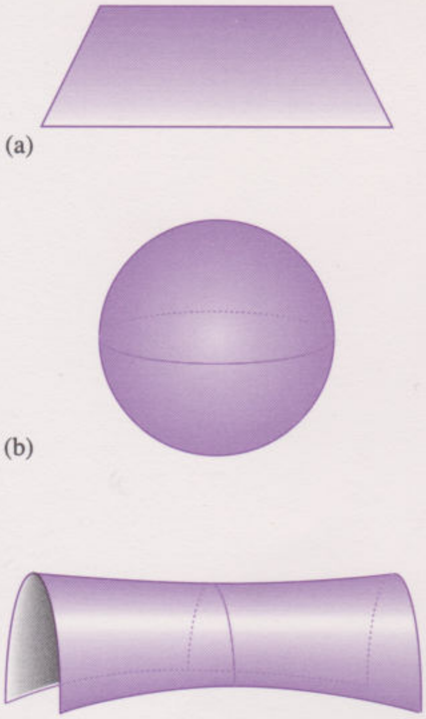
(a) (b) (c)

Figure 2.7 Analogues involving two-dimensional surfaces for the three possible types of overall geometry for a homogeneous and isotropic universe: (a) flat; (b) spherical; (c) hyperbolic.

2.5.1 Flat geometry

Flat geometry is the case for which normal, Euclidean geometry holds. This means, for example, straight lines can be extended to infinity, parallel straight lines do not intersect, the angles of a triangle add up to 180°, the circumference of a circle = $2\pi \times$ radius, etc.

This might be exactly the way it is with our Universe. The only important difference between the two-dimensional analogy and the real thing (apart from the extra spatial dimension, plus the time dimension) would be the fact that, whereas the sheet we have depicted in Figure 2.7a has an edge to it, the Universe would have no edge; it would extend to infinity. This being the case, it would also (unlike the sheet) have no central point.

2.5.2 Spherical geometry

On the surface of a sphere, the shortest distance between points (the equivalent of a straight line) is the arc of a great circle – that is to say, a circle drawn on the surface such that it lies in a plane passing through the centre of the sphere (see *Project file*, Subsection 1.2.3).

□ Do such 'straight lines' extend to infinity?

■ Having travelled along one of these straight lines for a distance equal to the circumference of the sphere, one is back where one started. So, no, they do not extend to infinity.

Two lines of longitude are examples of parallel 'straight lines' on a sphere, i.e. they are arcs of great circles.

□ Is it the case that these parallel lines do not intersect?

■ All lines of longitude meet at the poles, so they *do* intersect.

Imagine drawing a triangle of great circle arcs on a sphere. For example, two of the lines might be lines 0° and 90° of longitude, and the third one the equator (see Figure 2.8).

□ What can be said of the sum of the angles of a triangle drawn on the surface of a sphere?

■ The sum will be greater than 180°. In the example quoted, all three angles are 90°, making a total of 270°.

A circle can be drawn on a sphere by imagining a string stretched over its surface, pegged at one end with the pencil attached to the other. The length of the string is the 'radius' of the circle.

□ What can be said about the circumference of the circle in terms of its radius?

■ The circumference of a circle drawn on a sphere will be less than $2\pi \times$ radius.

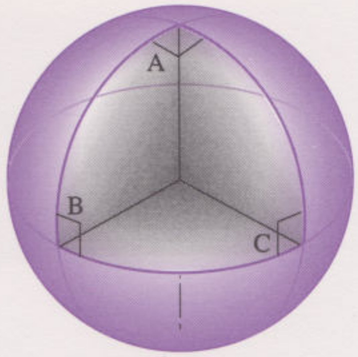

Figure 2.8 ABC constitutes a triangle drawn on a sphere.

It is easy to accept such conclusions regarding the geometry appropriate to the two-dimensional surface of a sphere (**spherical geometry**). What is more difficult to accept is the fact that *such conclusions might also apply to the geometry of three-dimensional space!* It all depends on the overall curvature of spacetime. Thus, in the real world, it is perfectly possible to imagine a vertical line starting off from the Earth's North Pole, continuing as a perfectly straight line out into space, but *not* going on for ever. After a certain distance it intersects with the Earth once more! It ends up at the South Pole. Thus, there is not only an obvious very short straight line connecting the poles passing *through* the Earth, there is this other one over the very large distance connecting the poles lying *outside* the Earth. So it is we can think of a perfectly good straight line in three-dimensional space, being in actuality a kind of 'arc of a great circle' as far as spacetime is concerned.

In such a spacetime, circles drawn in three-dimensional space would have circumferences less than $2\pi \times$ radius, the discrepancy being greater the larger the circle drawn. Also triangles would have angles adding up to $> 180°$. Again, this might not be noticeable for very small triangles, but the discrepancy would be there for triangles of cosmological dimensions – such as one constructed by taking sightings from three galaxies. If observers on each of the galaxies measured the angle subtended by the other two, and then compared results, they would find that the sum of these angles would be $> 180°$.

A universe of this kind would have a finite size; it would have a finite volume. And yet it would have no boundary. The astronaut would never come to a point where the Universe stopped. Travelling in a straight line, he/she would follow a closed path. Hence one calls such a universe a 'closed' universe. It would have no 'outside'. In this it is like the two-dimensional surface shaped as a sphere; it too has a finite size or area, but no matter how far one travels over its surface, one never encounters an edge.

2.5.3 Hyperbolic geometry

There is a third possibility – **hyperbolic geometry**.

□ Look at the saddle-shaped surface in Figure 2.7c. Draw your own conclusions regarding the properties of straight lines, parallel lines, the circumference of circles and the sum of the angles of triangles.

■ Straight lines would continue to infinity. Parallel lines would diverge rather than remain at the same distance from each other (the flat case), or intersect (the spherical case). This time the circumference of a circle would be $> 2\pi \times$ radius, and the angles of a triangle would add up to $< 180°$.

Again, this can be used as a two-dimensional analogy for the way the Universe might be. It is quite possible that circles and triangles drawn on a sufficiently large, cosmological scale would exhibit these features. As in the flat case, the Universe would extend to infinity in all directions.

In speaking of the Universe being infinite in extent, it is important to be clear what we mean. We mean that directly after the instant of the Big Bang the Universe was not only infinitely dense, but was *already* infinitely large. From $t = 0$ onwards, all distances within this infinite Universe expanded (so overall it became even more infinitely large!).

But often when speaking of 'the Universe' we do not mean this overall Universe. Rather we have in mind that particular part of it we observe today – what we have been calling 'the observable Universe'. This began infinitesimally small. Since then it has grown. You might think that this is the case because of the way *all* distances within the overall Universe are expanding. However, the size of today's observable Universe – radius 15 billion light years – is not so much governed by that as by the speed of light. As we saw in Chapter 1, Section 1.5, as time passes, light is able to reach us from ever more distant reaches of space, hence extending the horizon of observability. In effect the speed of light is outstripping the speed of expansion at the edge of the observable Universe revealing more and more of the overall Universe. We would expect this to continue until we are seeing so far out that the expansion speed asymptotically approaches that of light. Beyond that limit, the rest of the overall Universe remains hidden from us for ever.

Thus we see how important it is to draw a clear distinction between the 'overall Universe' and the 'observable Universe'. The latter is infinitesimally small at $t = 0$, and always remains finite in size – that size governed by the speed of light and the age of the Universe; the former is always finite if the geometry is spherical, and always infinite if the geometry is flat or hyperbolic – even from $t = 0$.

2.5.4 Which geometry for the Universe?

Having presented you now with the three possible geometries for our Universe, the question, of course, is which one is the winner?

Since there is as yet no direct evidence for the curvature of the Universe, it is difficult to decide. Tests involving the properties of triangles and circles show that Euclidean geometry seems fine. But that is what one would expect in any case if those geometrical figures are small. No matter what the curvature of the Universe might be, if one is looking at things on a very small scale, they will approximate to the flat case. No, what one would like to do is deal in circles and triangles with cosmological dimensions.

Until that becomes possible we must resort to other means. What general relativity says is that the overall curvature of the Universe is intimately bound up with the density of the matter causing that curvature. *It turns out that the critical density corresponds to the flat case.* A measurement of the density of matter is thus an alternative way of determining the overall geometry of the Universe.

In this work it is usual to define the **density parameter**, Ω, such that $\Omega =$ actual density/critical density. Thus, $\Omega = 1$ for the flat case. A higher density, such that $\Omega > 1$, corresponds to the spherical case, and $\Omega < 1$ to the hyperbolic case.

□ For the spherical case, what is the long-term future of the Universe?

■ With $\Omega > 1$, the density is greater than critical. As we saw in the last section, that implies that the Universe will one day halt its expansion and recollapse down to a Big Crunch.

Thus, the complete two-dimensional analogue of the Universe with $\Omega > 1$, taking account of the way it changes with time, is as follows: an infinitesimal dot (that represents the *whole* Universe) expands with time to reveal a spherical balloon; the expansion slows down to a halt; the balloon then deflates back to a dot.

☐ Describe the complete two-dimensional analogue of the *observable* Universe for the case with $\Omega < 1$.

■ An infinitesimal dot expands to reveal a surface that is saddle-shaped. This continues to expand for ever.

As we saw in Section 2.4, it is no easy matter deciding how the density actually compares with the critical value; we had to leave the question unresolved. So, not only is the long-term future of the Universe uncertain, we cannot be sure of the type of overall geometry it exhibits. However, we did hint in that last section that the inability to come up with a definitive inventory of the contents of the Universe might not be the end of the matter. Explaining what we meant by that is the subject of the next section.

Before ending this particular discussion, however, there is one more important point to be made; the effect of expansion on geometry. It should be emphasized that the expansion of the Universe cannot convert one type of geometry into another; whatever type of geometry the Universe exhibited at the beginning, that is the type of geometry it has today, and always will have.

This would be a good point at which to view video sequence 14, The geometry of the Universe. *Remember to read the associated notes first.*

Summary of Section 2.5 and SAQs

1 The 'observable' Universe is that part of the overall Universe such that light emitted from within it has had enough time since the Big Bang to reach us. With the passage of time, it encompasses an ever-growing fraction of the whole Universe.

2 The general theory of relativity replaces the notion of gravitational forces by the concept of warped spacetime.

3 It turns out that if the density is *critical* (expansion is halted only in the infinite future), the geometry is *always* flat and the overall Universe is infinitely large, even from time $t = 0$.

4 If the density is *greater than critical* (the Big Crunch scenario), the overall geometry of space is *always* spherical and closed – meaning that the overall Universe has at all times a finite volume, but has no boundary.

5 If it is *less than critical* (expansion for ever), the geometry is *always* hyperbolic, and the overall Universe is infinitely large, even from time $t = 0$.

SAQ 2.5 (Objective 2.2) In this section we have switched from talking about gravitational forces to the idea that spacetime is warped. Is this just a matter of personal preference? Could we just as easily have stuck with the Newtonian idea of gravitational forces? Discuss!

SAQ 2.6 (Objectives 2.2 and 2.3) Consider three-dimensional space. In Euclidean geometry the surface area of a sphere = $4\pi \times (\text{radius})^2$. If $\Omega > 1$, what could you say about the surface area of a sphere?

SAQ 2.7 (Objective 2.2) If the Universe were closed, so that it was modelled by the two-dimensional spherically-shaped surface, discuss whether or not some point in space would be the centre of this 'spherical' Universe.

2.6 The inflationary model

We have now presented you with the standard Big Bang theory of cosmology. Successful though it has been, it does raise a number of unanswered questions.

2.6.1 Homogeneity

Suppose we look at a region of space (call it region A) situated 15 billion light years away in a certain direction. We count the galaxies and estimate the density of matter in that region. Then we turn and look in the *opposite* direction. Again we measure the density of matter 15 billion light years away in a region B, i.e. 30 billion light years from A. What do we find? Assuming we are looking at regions sufficiently large that local variations associated with clusters and superclusters of galaxies can be ignored, the two densities are found to be the same. Why should this be so?

Our instinctive reaction is to say that when the Universe was much denser, the two regions had presumably been in contact with each other.

□ How could this solve the difficulty?

■ This would have given them a chance to equalize their densities, regardless of what they might have been originally. The densities would then have kept in step with each other as the Universe subsequently expanded, taking A and B to their respective positions today.

But there is a problem with this. Region A is at the edge of the observable Universe; the light we are receiving from it has taken the whole lifetime of the Universe, 15 billion years, to travel as far as the Earth (see Figure 2.9). Although region A was at the time – immediately after the Big Bang – very close to us, it has meanwhile been swept away from us at almost the speed of light. The light it emitted towards us directly after the Big Bang had little distance to go, but has had to swim against this tide of expanding space. It has only just arrived at the Earth. This is the first communication we have had with this region. But note: the Earth is only half-way between A and B; the light from A has a long way still to go before it reaches B for the first time. As we have seen, one of the consequences of relativity theory is that nothing can travel through space faster than light. It follows that there cannot have been any communication or contact between A and B. B is unaware of the existence of A – just as we on Earth are as yet unaware of galaxies at C that lie beyond A in Figure 2.9, but with which A is already in contact. So we are left with the question: why is the density of matter at A the same as at B?

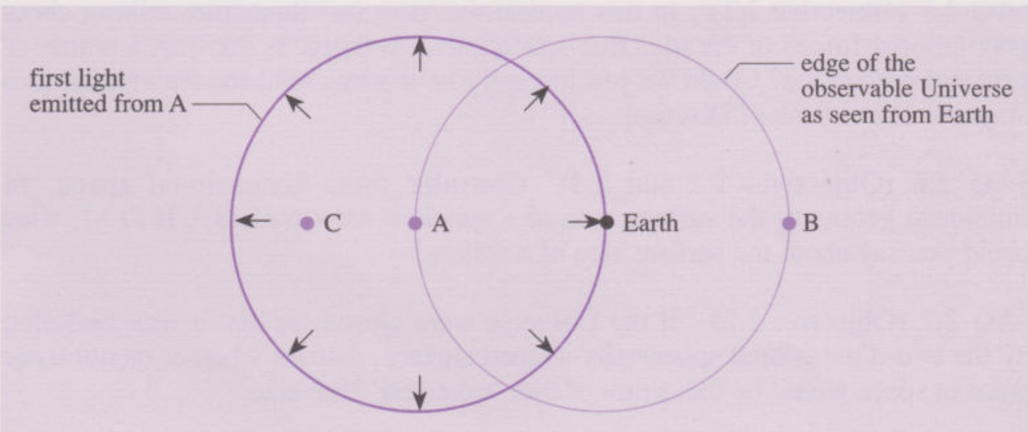

Figure 2.9 Light from region A has only just begun to arrive at Earth; it still has not reached region B.

2.6.2 Isotropy of the background radiation

The background microwave radiation has very nearly the same intensity and temperature regardless of direction. As we saw in Section 1.6, this is not immediately obvious from the data, because the Earth is in motion relative to the radiation and this introduces a Doppler shift correction. But when appropriate allowance is made for this effect, the intensity is found to be almost completely uniform. As we saw in Section 1.6, it is only at the level of 1 part in 100 000 that we begin to get the fluctuations recently discovered by the COBE group. It would seem, therefore, that the matter that originally gave out this light must have been in thermal equilibrium. But how could this be if we are dealing with radiation from sources on opposite sides of the Universe that, as in the preceding subsection, have never been in contact?

2.6.3 Flatness

The third outstanding problem is the fact, noted earlier, that the average density of the Universe is almost equal to the critical density. (Somewhere between 10% and 200% of the critical density qualifies as 'almost'!) We are essentially dealing with a universe of critical density. Why should this be a problem?

In order for the density parameter, Ω, to be close to 1 today (within a factor of 10), it had to be even closer to 1 in the past. This is because the deviation of Ω from 1 is expected to grow with time. If at any time $t = t_1$, Ω were *exactly* equal to 1, then at any later time, t_2, it would still be exactly equal to 1. But suppose it were less – for instance $\Omega = 0.9$, at t_1 – what would have happened to Ω by t_2? At first we might think it would remain at 0.9. After all, the total mass/energy in the Universe remains a constant throughout time, so if one were 10% short of mass at t_1, would that not also be the case at t_2? The answer is no. Because of the 10% deficiency, the expansion of the Universe between t_1 and t_2 would not have been subjected to the same degree of slowing down as would have been the case had $\Omega = 1$ originally. Therefore, the galaxies would be moving faster at t_2 than they would have been had $\Omega = 1$. Not only that, the Universe would be bigger than otherwise and the mutual gravitational attraction betweeen the galaxies correspondingly reduced. In order to make up for this, it would at this later stage be necessary to add *more* than 10% to the mass of the Universe in order eventually to bring the expansion to a halt. Thus, at t_2, Ω would be less than 0.9, and its deviation from unity would progressively increase with time. (It's a bit like the boy using his finger to block the growing hole in the dyke; if he delays, he will need more than his finger to do the trick.) In fact, the deviation from 1, in one model, grows at a rate proportional to $t^{2/3}$. Thus, for instance, whatever the deviation might have been 1 second after the Big Bang, 30 seconds later it would have already grown by a factor $30^{2/3}$; i.e. by a factor ≈ 10. The fact that today, billions of years later (i.e. 10^{17} seconds later), Ω still lies close to 1 (between 0.1 and 2), means that it must have been *incredibly* close to 1 soon after the Big Bang. Why should this have been the case?

2.6.4 Inflation

The easiest way out of these three types of difficulty – homogeneity, isotropy, and flatness – is to put them all down to initial conditions. We simply say that the Universe happened to start off from a very special state. Easy? Yes. But in the absence of any governing principle that would account for this special choice of initial conditions, such an explanation does not appear very satisfying.

The alternative is that, despite the arguments we have advanced to suggest that distant parts of the Universe could not have been in communication with each other, nevertheless, they *have* colluded together. Somehow or other, in the

very early stages of the Universe's development, they *were* in contact with each other, equalizing their density and temperature. Subsequent to that stage, they were carried apart by the expansion of space, at a rate greater than the speed of light (more of that later), to take up positions distant from each other. From these positions, they began the Hubble-type expansion we observe today.

How seriously we ought to take this idea depends crucially on our finding a mechanism that could give rise to the period of exceptionally rapid expansion – called **inflation**. Surprisingly, the most likely source of such a mechanism has been found, not through the study of astronomy and the large-scale structure of the Universe, but from the study of the behaviour of the Universe's smallest constituents!

The ultimate constituents of the cosmos are electrons, photons, and neutrinos, together with quarks. The last named are the parts from which neutrons and protons are built. (It takes three point-like quarks to make up either a neutron or a proton.) Other particles can be created in high-energy nuclear collisions, but these are short-lived, decaying quickly into the more familiar particles we find in nature. The particles interact with each other in various ways.

The four forces

One of the main tasks of elementary particle physicists is to examine the forces that give rise to these interactions. There are four:

(i) *Gravitational* This is the least effective force among elementary particles. This might strike you as surprising – given the importance of gravity in everyday life and in cosmology. But it is important to recognize that the masses of elementary particles are exceedingly small – so small that their mutual gravitational forces have so far proved too weak to be detected experimentally. Gravity assumes its importance for large-scale phenomena through the cumulative effect of vast numbers of these elementary particles.

(ii) *Weak nuclear* This force is responsible for the radioactive decay of many nuclei and for the interaction of neutrinos. Despite its name, its strength is typically 10^{33} times greater than that of gravity. The reason it figures so little in everyday life is that, unlike gravity which extends its influence indefinitely into space, the weak force has a very short range – 100 times smaller than the size of a nucleus. It is stronger than gravity *only* over this short distance scale.

(iii) *Electromagnetic* This includes both electric and magnetic forces and is responsible for holding electrons close to the nucleus in an atom. Over the short distances where the weak nuclear force operates, it is about 1 000 times stronger than the weak force. In addition, like gravity, it extends its influence to infinite distances. So, if it has the same range as gravity and is intrinsically so much stronger, why is it not the dominating force governing the overall structure of the Universe? Because, as you know, electric charge comes in positive and negative forms. Normally, matter has equal amounts of the two kinds and so the repulsive and attractive forces they exert tend to cancel out.

(iv) *Strong nuclear* This is the force that binds neutrons and protons together in the nucleus, and holds quarks together in the neutrons and protons. It is the strongest of all the forces – typically 100 times stronger even than the electromagnetic force. It binds protons together in nuclei despite their mutual electromagnetic repulsion. Again, its influence is not more generally felt because, like the weak nuclear force, its range is restricted – to about the size of the nucleus (10^{-15} m).

So, there we have it: four forces – all with vastly different strengths and ranges.

The unification of forces

But are they actually so different? Appearances can be deceptive! What one finds is that the strengths of the forces vary with the energy involved in the interaction. If particles collide at high energy, they react with different strengths to what they do at lower energy – hence the reason we prefaced our remarks about their relative strengths with words like 'typically' and 'about'. In fact, the numbers quoted above apply for particles colliding with kinetic energies at around 1 GeV (i.e. $\sim 10^{-10}$ J).

As elementary particle physicists probe upwards in energy, they begin to notice that the gap between the strengths of the forces narrows. Indeed, by the time an energy of 100 GeV ($\sim 10^{-8}$ J) is reached, the electromagnetic and weak forces are virtually indistinguishable. At this elevated energy it becomes clear that they are not separate forces at all; they are merely different manifestations of a single force – the newly named *electroweak* force. This was hailed as one of the great discoveries of the 1970s. It was an advance comparable to that of Maxwell in the 1860s when he recognized that electric and magnetic phenomena were merely different manifestations of a common reality – electromagnetism.

Can this process of unification be taken a step further? The belief is that it can. It is reckoned that at higher energies still, the strong force will weaken and the electroweak force will grow stronger. This will continue to the point where the two merge into a combined interaction called the *grand unified force*. The best estimate as to when this will happen points to an energy of between 10^{14} and 10^{15} GeV (i.e. around 10^4–10^5 J). The trouble is that this is a factor of about 10^{11} higher than anything achievable today in the laboratory! It is a colossal energy – the equivalent of the energy carried by a 5 kg shell travelling at 100 m s^{-1} – all concentrated into each subatomic particle. Fortunately, however, we may not *have* to achieve such energies. The theory does predict that there are likely to be low-energy side-effects. The fact that the magnitude of the electric charges on the proton and the electron are equal is thought to be one of them; the prediction that the proton might decay with a very long lifetime is a possibility that is currently being explored; and the theory gives predictions for the relative strengths of various weak and electromagnetic processes that can be tested at low energy.

More speculative still is the idea that at yet higher energies the grand unified force will merge with gravity. Thus, at exceedingly high energies there will be but one type of force.

Applying elementary particle physics to cosmology

Now, what is the relevance of all this for cosmology? Simply this: elementary particle physicists might have difficulty realizing in practice the kinds of energy where these unifications become manifest, but the Big Bang must surely have experienced these conditions – provided we are prepared to push our investigation of the Big Bang even closer to $t = 0$ than we did in Section 1.8, where you remember, we began the scenario at $t = 1$ s.

In fact, as we mentioned there, cosmologists these days venture to begin their account of the origins of the Universe at a time of 10^{-43} second! Obviously when they push things this far they are on less secure ground than when they are dealing with, say, the epoch of nucleosynthesis. And yet there is nothing we know that would prevent us applying the ideas of general relativity right back to that time. Now, you might think that if we are prepared to chance our arm to that degree, why not go the whole hog and start at $t = 0$. The trouble is that no-one knows how to marry together the ideas of quantum theory and general relativity. One is based on discontinuous point-like interactions subject to probability fluctuations; the other on the idea of smooth movement in a spacetime continuum. All we do know is that when one gets down to times of order 10^{-43} second, and distances equal to the size of the observable Universe at that sort of

time, the difficulties become horrendous. 10^{-43} s is, in fact, called the **Planck time**, so named after the German physicist Max Planck (1858–1947), the founder of quantum theory. Only subsequent to that time could general relativity as we know it begin to take charge.

At this stage, with the energy of particles about 10^{19} GeV (10^9 J) and a temperature of 10^{32} K, it is envisaged that the gravitational and grand unified forces have just assumed their separate identities. (Prior to this time there had just been the one force.) The particle mix will consist of quarks, electrons, neutrinos, etc. The quarks will be all mixed up together in a soup rather than being confined to groupings making up identifiable neutrons and protons. Not only that but the quarks and the electrons would be freely changing into each other. The expansion at this stage, like that taking place today, was governed by general relativity. Neighbouring regions of particles had the opportunity of reaching thermal equilibrium and homogenizing their density. By 10^{-35} s, the temperature had dropped to about 10^{28} K – the point where the particle energy has fallen to the critical value of 10^{15} GeV.

□ What is the significance of this energy?

■ This is when the character of the grand unified force undergoes a change: for the first time the strong nuclear and the electroweak forces separate out as distinctive forces with different strengths.

The immediate consequence of this is that the medium undergoes a dramatic transformation: there is a **phase change**. What do we mean by that?

Phase changes

The best-known examples of phase changes are when steam converts to water, and water to ice. The physical properties of the medium depend on the forces between the water molecules and the vigour of their motion. In the liquid phase, for example, the movements of the molecules are sufficiently vigorous that they can slide over each other and adopt different orientations relative to each other. But as the temperature drops, and the movements become less energetic, the attractive forces between the molecules can assume a greater measure of control. Below 0 °C (at normal atmospheric pressure), they can *fix* the positions of the molecules, and the liquid becomes a solid, i.e. ice.

This is not to say that the phase change occurs exactly at that temperature. There needs to be some triggering mechanism that determines where the ice crystal shall start to form – why it should nucleate from one location rather than another (a local temperature fluctuation, or the presence of a speck of dust perhaps?). In the absence of suitable nucleating centres, it is quite possible for the liquid to supercool – for its temperature to drop below 0 °C (as low as –20 °C) before crystallization occurs. When crystallization does occur from a deeply supercooled condition, the process is rapid – unlike the more familiar slow freezing that takes place on a pond, say. As the ice forms, heat is given out because the solid configuration represents a state of lower energy than the liquid phase. To convert ice back to liquid water, this energy would have to be restored; it is called the latent heat of fusion.

'False' and 'true' vacuums

It is believed that something of the same sort happened when the energies of the particles of the Universe dropped below 10^{15} GeV and the strengths of the forces between them began to change; the medium entered a supercooled condition, called the 'false vacuum'. 'False' because it is not the natural state to be in at that temperature, inasmuch as there is now a lower-energy state available; 'vacuum' because what we are concerned with here is intimately bound up with the properties of space itself.

As was the case with the formation of ice, there needed to be some triggering mechanism to initiate the phase change. This is thought to have been a quantum fluctuation of some kind. The phase change was one in which the false vacuum was transformed into the 'true vacuum' – the natural, lowest-energy condition for the space to be in. The transformation was dramatic.

To understand what happened, you need to know that one of the peculiar properties of the false vacuum was that it exerted a *negative* pressure. Instead of pushing outwards on other things, the tendency was to suck inwards – except, of course, there was nothing outside it to suck on. It is pressure *differences* that give rise to observable effects, not pressure itself. The negative nature of the pressure only manifested itself once pockets of true vacuum formed. True vacuum, behaving in the way one would expect a proper vacuum to behave – i.e. exerting zero pressure – found itself surrounded by the negative pressure of the false vacuum. It promptly expanded.

Superluminal expansion

And HOW it expanded! The normal type of expansion was temporarily suspended and an entirely different mechanism took over. Whereas previously the scale factor $R(t)$ in one model had been proportional to $t^{\frac{1}{2}}$, now it went exponential; it became proportional to $10^{\lambda t}$ where λ is a constant. Estimates show that, starting at about 10^{-35} s, the observable Universe doubled in size about every 10^{-34} s.

Recall that in the Einstein–de Sitter model $R(t) \propto t/t_0^{2/3}$. The Einstein–de Sitter model is a very particular model, even among those with critical density; for example, it has no radiation effects in it.

ITQ 2.4 By what factor would the observable Universe have grown between $t = 10^{-34}$ s and $t = 10^{-33}$ s?

The net result was that by the time the phase change was completed at about $t = 10^{-32}$ s, the diameter of the observable Universe had increased by an astonishing factor. How big? It is difficult to be sure, but we are talking of factors of the order of 10^{50} or possibly more! Thus the matter, having first had a chance to become homogeneous and to reach thermal equilibrium, was swept along by the tide of rapidly expanding space, at speeds exceeding that of light, to their respective starting positions for where the normal Hubble-type expansion recommenced.

In speaking of speeds in excess of that of light, we need to bear in mind that there is nothing amiss about this. True, Einstein's special theory of relativity imposes a limit on the speed of particles equal to the speed of light. But this speed limit only applies to motion *through* space. The movement caused by inflation (like that brought about by Hubble-type expansion) owes its origin to the expansion of space itself. When the movement of an object is brought about by that object being swept along by the tide of space, we are dealing with a totally different kind of situation. This type of movement is not subject to the same restriction. There is nothing to stop such speeds exceeding that of light. (Though you should note that a galaxy today receding from us with a speed in excess of that of light would lie beyond the bounds of our *observable* Universe, and so would not be seen by us.)

The early Big Bang scenario, therefore, looked something like that shown in Figure 2.10. Here we are able to fill in the gap between 10^{-43} s and the 1 s starting point we took for our description in Section 1.8.

See also Plate 3.42.

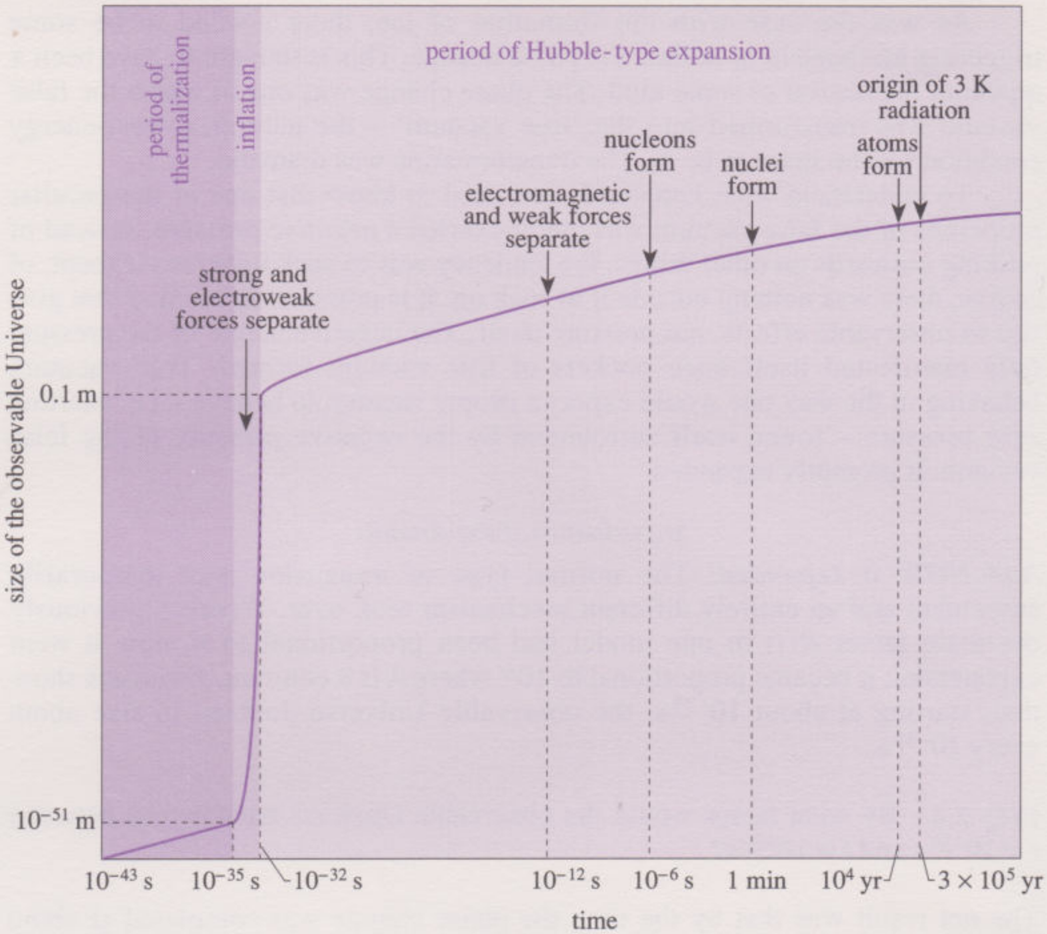

Figure 2.10 A summary of the various epochs of the Big Bang, undergone by the observable Universe (*not* drawn to scale).

The creation of matter and energy

When recalling the phase change from liquid water to ice, we mentioned how the drop to the lower energy state was acccompanied by a release of energy (the latent heat). The same thing happened with the early Universe. Energy was given out. This went towards maintaining the temperature of the mix to within a factor of about 10 of that which prevailed at the onset of inflation – about 10^{28} K. But more important still, it manifested itself in the creation of particles. Vast quantities of quarks and electrons were produced. Indeed, almost all the matter that makes up the Universe today owes its origins, not to whatever happened at time $t = 0$, but *to the release of energy accompanying inflation*. In fact, the matter that existed immediately prior to inflation must have been so severely diluted by the inflationary expansion, that it made up only a tiny proportion of the final mix.

Towards the critical density

A further, and most gratifying feature of this process for generating matter during inflation, is that it automatically generates the right amount to push the density towards the critical density! In other words, it was immaterial what the density of matter was prior to the onset of inflation; it was rapidly driven to a value that approached very, very closely that which was necessary just to halt the subsequent Hubble-type expansion. It can be shown, for example, that if inflation led to a change in size by a factor of 10^{50}, the deviation of Ω from 1 would have been reduced by a factor 10^{100}. Thus, inflation leads in a natural way to the production of a universe that conforms closely to a critical density model. It does not convert the Universe to one that conforms *exactly* to a critical density model;

if the density before the onset of inflation was greater (less) than critical, it will still be greater (less) afterwards, though only barely so.

In terms more closely allied to the geometrical type of thinking used in general relativity, we might imagine a universe with a density greater than critical being represented by a two-dimensional rubber surface in the shape of a balloon. As inflation proceeds, the balloon is blown up to an immense size, such that, to all intents and purposes, a small area on the surface of the balloon appears to be flat (see Figure 2.11). Our *observable* Universe would occupy just such a tiny flat area. Note that the flatness of the small area does not depend on what the curvature of the balloon was before it was inflated; it could have had any curvature you like, provided that the degree of inflation was great enough.

□ Supposing the density were *less* than critical, what would be the two-dimensional analogue for the Universe undergoing inflation?

■ It would be a saddle-shaped rubber sheet which, after inflation, had been almost completely flattened over the tiny part constituting the 'observable' Universe.

At a stroke, the inflation idea solves the homogeneity, isotropy, *and* flatness problems that afflicted our earlier model of the Big Bang. Indeed, it might even be that inflation holds the key to understanding the slight inhomogeneities that do occur – those responsible for triggering the formation of galaxy clusters and those that manifest themselves as the ripples in the microwave background radiation. The inflation process is the product essentially of a quantum process. Thus, it is claimed, the fluctuations one sees today on the cosmic scale might be the original quantum fluctuations writ large. Whether this claim is justified is not clear at the time of writing. Even so, the undoubted successes of inflation theory are considerable. It reflects great credit on its originators, especially Alan Guth (1947–), the American cosmologist who contributed important ideas in 1980.

Phase boundaries

Not that the original formulation was without problems. One particular difficulty was that the initial proposal envisaged not one but many regions of true vacuum forming simultaneously out of the false vacuum. As they grew, they would have met up with each other at boundaries called **phase boundaries**. This would be the equivalent of several ice crystals starting to form in water all at once. They give rise to discontinuities where their surfaces meet up. These boundaries are called 'defects'; the crystals cannot join up to form one big crystal because their symmetry axes lie in different directions. When each individual crystal originally began to form, it had to adopt certain directions in space for its axes of symmetry. There was nothing to distinguish one direction from another – they were all equally as good as each other – so which way it went was all a matter of chance. Thus, crystals that start growing independently of each other almost certainly end up pointing in different directions.

So it would be in the early Universe. Different regions of true vacuum would start to grow. But there are a variety of ways in which the strong and the electroweak forces can separate out from each other. These differing ways lead to differing values for the strengths of the forces and for the masses of the quarks and electrons. It becomes a matter of chance which particular set of allowed values any particular region gets. Thus, there is no coordination between the different regions of true vacuum, so that, when they meet up with each other at a boundary, the physics is different either side of the boundary. As you might anticipate, this leads to trouble. In point of fact, the theory shows that at the boundaries there should be several noticeable effects, depending on the particular type of intersection between regions. Thus, one would expect two-dimensional 'walls' (domain walls). These the theory shows would be characterized by large

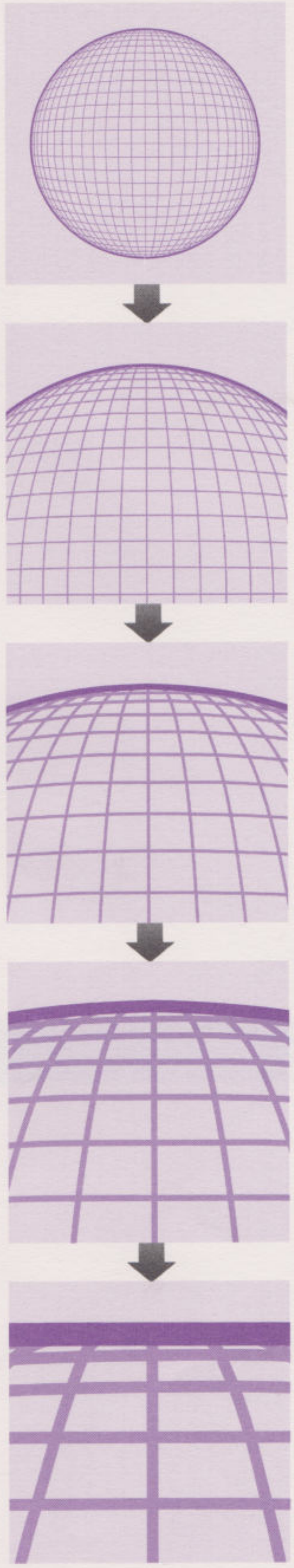

Figure 2.11 A spherical balloon analogue for illustrating how the inflation of the Universe at an early epoch would have resulted in a flat geometry for the observable Universe regardless of what the curvature might have been prior to inflation.

concentrations of energy. There could be other features: one-dimensional cosmic strings with mass 10^{18} kg per metre length, and point-like magnetic monopoles of mass 10^{16} times that of the proton. Moreover, the regions should have been sufficiently small that the observable Universe today should encompass large numbers of them. The observation of boundary defects should be fairly commonplace – which they are not, none having been detected, so far at least. It would also presumably be quite easy to detect regions of space with different physical constants; their emitted spectra would fail to match those found in our own laboratories.

This problem was subsequently overcome by later workers. They suggested a somewhat different kind of quantum or thermal fluctuation for triggering the initial growth of a region of true vacuum. This led to the expectation that each region would end up much, much larger than Guth had originally envisaged. Each region would have been such that, prior to inflation, it was contained within a light horizon – meaning that a light ray would have had sufficient time to cross from one side of the region to the other in the time then available since $t = 0$. In other words, it was a region throughout which thermal equilibrium and a homogeneous density could have been established. How big was this region? Assuming the onset of inflation occurred at about 10^{-35} s, light could have travelled $3 \times 10^8 \times 10^{-35}$ m $\approx 10^{-27}$ to 10^{-26} m. This region was then expanded by a factor of about 10^{50} during inflation to a size of about 10^{24} m afterwards. That then was the size, immediately after inflation, of the region of the Universe having the same values of the physical constants, and additionally, being in thermal equilibrium and being of homogeneous density. We must compare that distance with what was *then* the size of today's observable Universe. Towards the end of the inflationary period, this was just 10 cm across. In other words, our observable Universe today is an absolutely minute corner of an unbelievably vast region, all with the same values of quark and electron masses and strengths of forces. Typically the distance between ourselves and the nearest boundary defect separating our region of true vacuum from the next might be some 10^{25} times the radius of the observable Universe. And goodness knows how many other regions lie beyond there! What a salutary thought *that* is!

Summary of Section 2.6 and SAQs

1. The inflationary scenario has been proposed to solve the problems of the homogeneity of matter, the isotropy of the background radiation, and the flatness of space.
2. It proposes a very brief period of exceptionally rapid exponential expansion in the early Universe.
3. A study of the behaviour of subatomic particles at high energy suggests that inflation was caused by a phase change resulting from the 'freezing out' of the strong nuclear and the electroweak forces at a temperature of 10^{28} K. At earlier epochs and higher temperatures, these forces had been indistinguishable components of a common force – the grand unified force.
4. Almost all the matter and energy in the Universe today was created *during* this period of inflation.

Note that TV programme 8, Cosmology on trial, *provides an assessment of the evidence and the arguments upon which our modern ideas of cosmology have been built.*

SAQ 2.8 (Objective 2.6) Describe in your own words the physical changes that accompany the phase change from steam to liquid water.

SAQ 2.9 (Objective 2.6) During inflation, when distances were doubling in size every 10^{-34} s, how long would it have taken for the observable Universe to grow from a diameter 10^{-26} m to 10^{-23} m?

SAQ 2.10 (Objective 2.6) What was the average speed of the matter at the edge of the observable Universe during an interval of 10^{-34} s, towards the end of the inflation period, if its size at the end of the interval was 0.1 m? Compare your estimate with the speed of light.

SAQ 2.11 (Objective 2.6) How does Hubble-type expansion differ from inflation in respect of the temporal behaviour of Ω?

2.7 Postscript: the Universe that amounts to nothing

The last section was probably as good a point as any to end our Course. But just in case you are feeling a bit cowed by what you have now learnt of the sheer immensity of the Universe (and who isn't?), here we offer a few final words of comfort.

All the matter and energy created during the inflationary period – out of vacuum – is mind-boggling. But let us not overlook the fact that energy comes in both positive and negative forms. Heat, electromagnetic radiation, kinetic energy of moving bodies, and the locked-up relativistic energy in mass (mc^2) are manifestations of positive energy. The gravitational attraction between masses constitutes negative energy, if the zero of gravitational energy is set at infinite separation. The more matter that was created during inflation the more that matter contributed positive energy; but it *also* increased the gravitational negative energy. When the sums are carried out, a good case can be made for saying that, in a certain sense, all the positive energy in the cosmos is exactly cancelled out by all the negative gravitational energy. Not only that, there is as much positive electric charge in the world as negative – again leading to exact cancellation. And yet further, we note that with as many heavenly bodies rotating one way as another, the net angular momentum in the Universe is zero. And the net momentum of the Universe is zero, etc. Indeed, it appears that in the ultimate analysis, the Universe adds up to precisely *nothing*. Just think: you have spent the past few months studying the nature, the origin, and the future of absolutely nothing – albeit an ingenious rearrangement of nothing! The next time you look up at the stars and are made to feel small, you might like to bear that in mind.

Objectives for Chapter 2

After studying Chapter 2 (and any associated audio, video or TV material), you should be able to:

2.1 Give brief definitions of the terms, concepts and principles listed at the end of the Objectives.

2.2 Explain how the general theory of relativity provides a framework for discussing cosmology in terms of the expansion and curvature of spacetime,

in particular relating these to the finite or infinite extent of the Universe, and the lack of a boundary to the Universe.

2.3 Describe what is meant by spacetime and, given the appropriate formulae, perform simple calculations involving space and time.

2.4 Discuss the various possibilities for the future development of the Universe and how these relate to the critical density.

2.5 Discuss the measured value of the density of the Universe in relation to the critical value, including the possibility of there being dark matter.

2.6 Outline the inflationary model of the early Universe, explaining how this can solve the problems posed by the homogeneity of matter, the isotropy of the microwave radiation, and the observation that the density of the Universe appears to be close to the critical value.

List of scientific terms, concepts and principles used in Chapter 2

Term	Page	Term	Page
Big Bounce	44	inflation	58
Big Crunch	44	open universe	44
closed universe	44	phase boundary	63
dark matter	46	phase change	60
density parameter, Ω	54	Planck time	60
flat geometry	52	spacetime	41
Heat Death	46	spherical geometry	53
hyperbolic geometry	53		

ITQ answers and comments for Chapter 1

ITQ 1.1

The observed frequency, f_{obs}, of a Doppler-shifted signal is related to the emitted frequency, f_{em}, by Equation 2.3 in Book 1:

$$v = c \times \left| f_{\text{em}} - f_{\text{obs}} \right| / f_{\text{obs}}$$

(where we have dropped the subscript r for 'radial', and made the speed equal to the speed of light c).

The frequency is related to the wavelength by $c = f\lambda$ (Equation 1.2 in Book 1) and so rewriting the Doppler equation in terms of wavelengths gives

$$v = c \times \left| \frac{c}{\lambda_{\text{em}}} - \frac{c}{\lambda_{\text{obs}}} \right| \frac{\lambda_{\text{obs}}}{c}$$

Rearranging this formula gives

$$\frac{v}{c} = \frac{\left(\lambda_{\text{obs}} - \lambda_{\text{em}}\right)}{\lambda_{\text{em}}}$$

where we have dropped the modulus sign, $||$, because $\lambda_{\text{obs}} > \lambda_{\text{em}}$ here.

The right-hand side of this formula is the redshift, z (Equation 1.2). Hence

$$z = v/c$$

ITQ 1.2

(a) According to Hubble's law, $z = H_0\, d/c$. So the distance to a quasar with a redshift, $z = 0.12$ is

$$\begin{aligned} d &= (z \times c)/H_0 \\ &= (0.12 \times 3 \times 10^5\,\text{km}\,\text{s}^{-1})/(75\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}) \\ &= 480\,\text{Mpc} \end{aligned}$$

(b) Since 0.12 is (just about) a small redshift we adopt the Newtonian Doppler-shift interpretation and write $z = v/c$. Hence

$$\begin{aligned} v &= 0.12 \times 3 \times 10^5\,\text{km}\,\text{s}^{-1} \\ &= 3.6 \times 10^4\,\text{km}\,\text{s}^{-1} \end{aligned}$$

So the distance is increasing at $36\,000\,\text{km}\,\text{s}^{-1}$.

ITQ 1.3

(a) The age of the Universe is assumed to be given by

$$t_0 = 1/H_0$$

Using a value of $H_0 = 50\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}$

$$t_0 = \frac{1}{50\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}} = \frac{1}{50}\,\text{s}\,\frac{\text{Mpc}}{\text{km}}$$

Now $1\,\text{Mpc} = 3.1 \times 10^{19}\,\text{km}$, so $(\text{Mpc}/\text{km}) = 3.1 \times 10^{19}$, and

$$\begin{aligned} t_0 &= \frac{1}{50} \times (3.1 \times 10^{19})\,\text{s} \\ &= 6.2 \times 10^{17}\,\text{s} \\ &= 2.0 \times 10^{10}\,\text{years} \end{aligned}$$

(b) If we double H_0 we *halve* the age, and so, for $H_0 = 100\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}$

$$t_0 = 1.0 \times 10^{10}\,\text{years}$$

ITQ 1.4

The mass difference

$$m_{\text{n}} - m_{\text{p}} = 2.30 \times 10^{-30}\,\text{kg}$$

The rest energy is mc^2, so the difference in rest energy

$$\begin{aligned} e_{\text{n}} - e_{\text{p}} &= (2.30 \times 10^{-30}\,\text{kg}) \times (3.00 \times 10^8\,\text{m}\,\text{s}^{-1})^2 \\ &= 2.07 \times 10^{-13}\,\text{J} \end{aligned}$$

Equating the difference to a *thermal energy*, kT, gives a corresponding temperature of

$$\begin{aligned} T &= (2.07 \times 10^{-13}\,\text{J})/(1.38 \times 10^{-23}\,\text{J}\,\text{K}^{-1}) \\ &= 1.50 \times 10^{10}\,\text{K} \end{aligned}$$

ITQ 1.5

Substituting $T = 10^{10}\,\text{K}$ into Equation 1.22 gives a neutron to proton number ratio of

$$\begin{aligned} \frac{n}{p} &= 10^{-\left(6.5\times10^9\,\text{K}\right)/T} \\ &= 10^{-\left(6.5\times10^9\,\text{K}\right)/\left(10^{10}\,\text{K}\right)} \\ &= 10^{-0.65} \\ &= 0.22 \end{aligned}$$

i.e. a neutron : proton ratio of approximately 1 : 5.

[*Comment*: Equation 1.22 only applies to reactions 1.21a and 1.21b. Some neutrons are also lost via reaction 1.23, so the ratio of neutrons to protons will be slightly lower.]

ITQ answers and comments for Chapter 2

ITQ 2.1

(a) $20\,\text{mm}\,\text{s}^{-1}$; (b) $60\,\text{mm}\,\text{s}^{-1}$.

The rate of expansion of the rubber is uniform everywhere, so if a distance of 50 mm expands at the rate of $10\,\text{mm}\,\text{s}^{-1}$, twice the distance, 100 mm, will expand by twice the amount, namely, $20\,\text{mm}\,\text{s}^{-1}$. Similarly, six times the distance, 300 mm, will expand at six times the rate, i.e. $60\,\text{mm}\,\text{s}^{-1}$.

ITQ 2.2

From Equation 2.1

$$\begin{aligned} t_\text{a} &= t_\text{c}(1-0.9^2)^{\frac{1}{2}} = t_c(1-0.81)^{\frac{1}{2}} \\ &= t_\text{c}(0.19)^{\frac{1}{2}} \end{aligned}$$

therefore

$$t_\text{a}/t_\text{c} = 0.44$$

ITQ 2.3

If the mass of the neutrino, m_ν, is $2 \times 10^{-5}\, m_\text{e}$, and m_e is $\approx 1/2\,000\, m_\text{p}$, then

$$m_\nu \approx 10^{-8}\, m_\text{p}$$

The proton mass is nearly the same as the neutron mass. So, with 10^8 neutrinos for each baryon, that means the mass of the neutrinos would equal that of the baryons, i.e. it could make up to a few per cent of the critical density.

ITQ 2.4

The time elapsed is $10^{-33}\,\text{s} - 10^{-34}\,\text{s} = 9 \times 10^{-34}\,\text{s}$. Therefore, the size increases by a factor of $2^9 = 512$.

SAQ answers and comments for Chapter 1

SAQ 1.1

From Equation 1.2 we have

$$\begin{aligned} z\lambda_\text{em} &= \lambda_\text{obs} - \lambda_\text{em} \\ \text{thus } \lambda_\text{obs} &= \lambda_\text{em}(1+z) \end{aligned}$$

Substituting for $\lambda_\text{em} = 121.6\,\text{nm}$,

$$\begin{aligned} \lambda_\text{obs} &= 121.6(1+0.12)\,\text{nm} = 121.6 \times 1.12\,\text{nm} \\ &= 136.2\,\text{nm} \end{aligned}$$

This is also in the ultraviolet part of the spectrum.

SAQ 1.2

With *z small*, we use Equation 1.4, $v = H_0 d$

$$\begin{aligned} \text{Thus } v &= 75\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1} \times 350\,\text{Mpc} \\ &= 26\,250\,\text{km}\,\text{s}^{-1} \end{aligned}$$

Therefore

$$\begin{aligned} v/c &= \frac{26\,250\,\text{km}\,\text{s}^{-1}}{300\,000\,\text{km}\,\text{s}^{-1}} \\ &= 0.088 \end{aligned}$$

Because v is much smaller than c, then $z = v/c$ (Equation 1.3),

$$\text{so } z = 0.088$$

Thus z is indeed small, so the calculation is self-consistent.

SAQ 1.3

Bearing in mind that $1\,\text{Mpc} = 3.1 \times 10^{19}\,\text{km}$, a value of H_0 of $50\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}$ corresponds to

$$\frac{50}{3.1 \times 10^{19}}\,\text{s}^{-1}$$

$$= 1.61 \times 10^{-18}\,\text{s}^{-1} \text{ (see also ITQ 1.3)}$$

The age of the Universe is thus

$$\begin{aligned} t_0 &= 2/(3H_0) = 4.1 \times 10^{17}\,\text{s} \\ &= 1.3 \times 10^{10} \text{ years} \end{aligned}$$

Similarly, if the value of H_0 is assumed to be $100\,\text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1}$, the age becomes $2.1 \times 10^{17}\,\text{s} = 6.6 \times 10^9$ years. The possible age of the Universe ranges from 6.6×10^9 years to 1.3×10^{10} years.

SAQ 1.4

At each instant of cosmic time it is possible to draw a graph of distance versus recessional speed for the *nearby* galaxies. It yields a straight line relationship such that the *constant* of proportionality is Hubble's constant. In this sense it *is* a constant. The confusion comes from the fact that a different plot of distance versus recessional speed applies to each and every instant of cosmic time. The various straight lines have different slopes, and hence different values for Hubble's constant.

[*Comment*: The nearby galaxies (to which we can apply the recessional speed idea) are sufficiently close that they do give us a 'snapshot' at an instant of cosmic time – deceleration does not show up because we are not looking back far enough in time.]

SAQ 1.5

Just as Equation 1.17 gives a value of the recessional speed that is just sufficient to allow a galaxy to escape the restraining pull of the other galaxies within the sphere centred on our position at the Earth, so the same kind of formula can be applied to the escape speed of an object

from the surface of the Earth. [*Comment*: This was a topic in Section 1.5 of *Preparatory science*.]

SAQ 1.6

Converting from Mpc to km, as we did in the answer to SAQ 1.3, the first of the two values of H_0 to be considered becomes $1.61 \times 10^{-18}\,\mathrm{s}^{-1}$.

From Equation 1.18

$$\begin{aligned}\rho &= 3H_0{}^2/(8\pi G)\\ &= \frac{3\times(1.61\times10^{-18}\,\mathrm{s}^{-1})^2}{8\pi\times6.67\times10^{-11}\,\mathrm{N\,m^2\,kg^{-2}}}\\ &= 4.6\times10^{-27}\,\mathrm{kg\,m^{-3}}\quad(1\,\mathrm{N}\equiv1\,\mathrm{kg\,m\,s^{-2}})\end{aligned}$$

Similarly, for $H_0 = 100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$

$$\rho = 1.9\times10^{-26}\,\mathrm{kg\,m^{-3}}$$

With the mass of a hydrogen atom = 1.67×10^{-27} kg, we see that in the first case there are just under 3 atoms per m^3, and in the second just over 10.

SAQ 1.7

The shape of a black-body spectrum certainly depends on the temperature of the source. But the height of the curve also depends on the size and distance of the source, and upon interstellar extinction. The measurement of spectral flux density at a single wavelength, therefore, cannot by itself allow one to determine which spectrum one is sampling.

SAQ 1.8

We see from Figure 1.11 that, on the assumption that the density of baryonic matter today is between $0.05\,\rho_c$ and $0.01\,\rho_c$, there was virtually no production of heavy nuclei in the Big Bang. We must, therefore conclude that all the heavy elements we see about us today must have originated in the fusion reactions occurring in stars.

SAQ answers and comments for Chapter 2

SAQ 2.1

With the rubber balloon analogy, all the rubber began with a very small size. But one cannot look at the *surface* of the expanded balloon and ask 'At which point on this surface was all the rubber originally concentrated?' Similarly, we cannot today look at the expanded state of space and locate some special point in it where all the space once was. All points in space have equal status to all other points, just as all points on the balloon have equal status. Just as it does not make sense to think of the centre of the surface of the balloon lying somewhere in its two-dimensional surface, so it does not make sense to think of the centre of the Universe lying somewhere in its three-dimensional space.

SAQ 2.2

(a) and (b): The wavelength would be shorter than would have been the case without this motion relative to the cluster. What we are dealing with here is a normal type of Doppler shift, similar to that which occurs when *any* light source moves *through* space towards us. As far as we ourselves are concerned, the source's motion affects the wavelength of the light both at emission and at receipt.

[*Comment*: How does that tie up with what we said earlier about the recessional speed of a galaxy *not* affecting the wavelength of the light as it is emitted? By this we meant the speed of recession characteristic of that location in space *due to the expansion of the Universe*. This is much more likely to be exhibited by the overall motion of a cluster of galaxies than by any individual galaxy. The motion of a particular galaxy will then be a superposition of the cosmological movement (which Hubble-shifts the wavelength at receipt, but not at emission), and 'normal' movement through space relative to the cluster – the radial component of the latter gives rise to a Doppler shift, which affects the wavelength both at emission and receipt. In Book 3, Chapter 4, these two motions were called *Hubble flow* and *peculiar motion* respectively.]

SAQ 2.3

According to ITQ 2.2, the astronaut's time t_a is related to the controller's time t_c via

$$t_a = 0.44\,t_c$$

Moreover, for the astronaut

$$\begin{aligned}x^2+y^2+z^2 &= 0\\ \text{therefore } s^2 &= c^2\,(0.44)^2 t_c{}^2 - 0\\ \text{and } s &= 0.44\,ct_c\end{aligned}$$

According to the controller, the distance to the planet, $(x^2 + y^2 + z^2)^{1/2}$, is

$$\begin{aligned}&= \text{speed}\times\text{time}\\ &= 0.9\,c\times t_c\\ \text{therefore } s^2 &= c^2t_c{}^2 - 0.9^2\,c^2t_c{}^2\\ &= c^2t_c{}^2(1-0.9^2) = 0.19\,c^2t_c{}^2\\ \text{and } s &= 0.44\,ct_c\end{aligned}$$

This is the same as for the astronaut.

SAQ 2.4

As the galaxies begin to come towards us, the light they give out would be blueshifted rather than redshifted. However, it takes time for the light to reach us, so it would be the nearer galaxies that would show the blueshift first.

SAQ 2.5

No. In the first place, general relativity provides a natural explanation of the identical motions of objects close to 'gravitating' bodies; the old Newtonian idea of gravitating forces was faced with the unexplained coincidence of why gravity always managed to pull on heavy objects with just the right amount of extra force to get them to travel along the same paths as lighter objects. But more importantly, general relativity provides more accurate predictions of motions, for example, the motion of Mercury, or the bending of light. It also predicts that time is affected close to massive bodies, as experimentally observed. With all these successes we are led not only to accept general relativity as our best understanding of what happens on the local scale, but also what it implies about the overall structure of the Universe.

SAQ 2.6

For $\Omega > 1$ (spherical geometry), the circumference of a circle is less than $2\pi \times$ radius, so we would expect that the surface area of a sphere would be less than $4\pi \times (\text{radius})^2$.

SAQ 2.7

Whereas it is true that an ordinary sphere has a centre, that centre is not to be found *in* the two-dimensional surface. It is this surface that is the analogy for our three-dimensional space, and the surface has no centre. So there will be no point in the three-dimensional space of the Universe that can be regarded as its centre. [*Comment*: It does not matter what the geometry of the Universe is; all points in space are on an equal footing; there is no centre to the Universe.]

SAQ 2.8

In steam, the motion of the molecules is so vigorous that the intermolecular forces are not able to hold the particles together and the medium behaves as a gas. But with the reduction in average speed that accompanies the drop of temperature below 100 °C at normal atmospheric pressure, the intermolecular forces can hold the molecules together, though not at fixed positions relative to each other. This is characteristic of the liquid phase.

The transformation from the steam phase to the liquid phase does not necessarily occur immediately the temperature drops below 100 °C. For water droplets to form, there must be suitable nucleating centres (otherwise there would be no reason why the droplet should start forming at one location rather than another). In the absence of such centres, the temperature of the steam can drop well below 100 °C before any phase change occurs. [*Comment*: In this case it is said to be supercooled.]

When condensation does take place, there is a release of energy – energy that must be restored if the liquid is subsequently to be vaporized again. [*Comment*: This heat is known as the latent heat of vaporization.]

SAQ 2.9

The increase in size is by a factor

$$10^{-23}/10^{-26} = 10^3 = 1\,000$$

So the question becomes: How many times do we have to multiply 2 by itself to get 1 000? By trial and error we note that $2^5 = 2 \times 2 \times 2 \times 2 \times 2 = 32$, and $32 \times 32 \approx 1\,000$. So the answer is that we must multiply 2 by itself 10 times.

(If you are familiar with the use of logarithms, you may have used the following method:

$$\begin{aligned} 2^x &= 10^3 \\ \text{so } x \log_{10} 2 &= 3 \\ \text{therefore } x &= 3/0.301 \approx 10) \end{aligned}$$

Thus, the Universe doubled in size 10 times over. This took $10 \times 10^{-34}\,\text{s} = 10^{-33}\,\text{s}$.

SAQ 2.10

During the last 10^{-34} s, it would have doubled its size. If its diameter was 0.1 m at the end of this period, it must have been 0.05 m at the commencement of this period. Thus, it covered a distance of 0.05 m to reach 0.1 m in 10^{-34} s.

If v is the speed of the outermost regions of the observable Universe,

$$\begin{aligned} v &= 0.05\,\text{m}/10^{-34}\,\text{s} \\ &= 0.05 \times 10^{34}\,\text{m s}^{-1} \\ &= 5 \times 10^{32}\,\text{m s}^{-1} \end{aligned}$$

Compared to $3 \times 10^8\,\text{m s}^{-1}$ for the speed of light, this is a factor of about 10^{24} times higher. Quite fast!

SAQ 2.11

In Hubble-type expansion, any initial deviation of Ω from 1 is progressively accentuated with time. On the other hand, with the inflationary type of expansion, the opposite occurs: Ω is driven ever closer to 1

Acknowledgements

Grateful acknowledgement is made to the following sources for permission to reproduce material in this book:
Figure 1.10 courtesy of NASA/CSFC and COBE Science Working Group; *Figure 1.11* T.P. Walker *et al.* (1991) Primordial nucleosynthesis redux, *The Astrophysical Journal*, vol. 376, no. 1, Part 1, p. 65 © 1991 by the American Astronomical Society; all rights reserved.

Index